




3 1761 10374374 6



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743746>

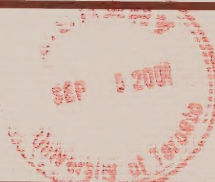
12-001



252

Government
Publications

SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2001

•

VOLUME 27

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2001 • VOLUME 27 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2001

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

August 2001

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
E. Rancourt (Production Manager)
C. Patrick

R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
D. Holt, *University of Southampton, U.K.*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*

G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistics Canada*
D. Pfeiffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *The Urban Institute*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

SURVEY METHODOLOGY
A Journal Published by Statistics Canada
Volume 27, Number 1, June 2001

CONTENTS

In This Issue	1
---------------------	---

Waksberg Invited Paper Series

G. NATHAN Telesurvey Methodologies for Household Surveys – A Review and Some Thoughts for the Future?	7
--	---

Special Section on Composite Estimation

A.C. SINGH, B. KENNEDY and S. WU Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design	33
W.A. FULLER and J.N.K. RAO A Regression Composite Estimator with Application to the Canadian Labour Force Survey	45
P. BELL Comparison of Alternative Labour Force Survey Estimators	53
J. GAMBINO, B. KENNEDY and M.P. SINGH Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation	65

Regular Papers

J.-K. KIM Variance Estimation After Imputation	75
T.E. RAGHUNATHAN, J.M. LEPKOWSKI, J. VAN HOEWYK and P. SOLENBERGER A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models	85
J. DUFOUR, F. GAGNON, Y. MORIN, M. RENAUD and C.-E. SÄRNDAL A Better Understanding of Weight Transformation Through a Measure of Change	97
P. ARDILLY and D. LE BLANC Sampling and Weighting a Survey of Homeless Persons: A French Example	109

In This Issue

This issue of *Survey Methodology* contains the first in an annual invited paper series in honour of Joseph Waksberg. A brief description of the newly instituted series and a biography of Joseph Waksberg are given before the paper itself. I would like to thank Danny Levine for writing the biography of Joseph Waksberg. I would also like to thank David Binder, Paul Biemer, Graham Kalton, and Chris Skinner, the current members of the Committee for choosing a very prominent survey researcher to author the first paper of the Waksberg Invited Paper Series. My special thanks are due to Graham Kalton who, as the founding Chairman of the Committee, took the lead, negotiated the necessary arrangements with Westat, the *American Statistical Association* and *Survey Methodology* to set the wheel in motion and worked hard to meet the deadline set by the journal for publication of the June Issue.

The author of the Waksberg Invited Paper for 2001 is Gad Nathan. His paper, "Telesurvey Methodologies for Household Surveys – A Review and Some Thoughts for the Future", presents a methodological history of telephone surveys from the 1930s up to the present day. Topics covered include sampling designs, sampling frames, coverage, nonresponse and weighting. He finishes the paper by describing some of the challenges and opportunities posed by more recent developments such as email, the internet, cell phones, and other emerging technological and social changes.

This issue of *Survey Methodology* also includes a special section on composite estimation with four papers. The first of these papers, by A.C. Singh, Kennedy and Wu, describes the method of regression composite estimation developed by Singh and colleagues over the past few years. They compare the new approach to previous methods of composite estimation, most notably the K -composite and the AK -composite estimators. The paper also includes a heuristic description and motivation of the new approach. Advantages of the new approach are that it yields a single set of estimation weights, leading to internal consistency of estimates, while improving on the efficiency of conventional regression estimators.

Fuller and Rao give an analytical evaluation of the properties of regression composite estimation. They first describe two earlier variants of regression composite estimation called modified regression estimators ($MR1$ and $MR2$), and analyse the efficiency and behaviour of the estimates over time using a simple time series model for the survey panel estimates. They conclude that a modification which can be viewed as a compromise between $MR1$ and $MR2$ would have the best properties overall.

In his paper, Bell compares a range of alternative estimators for use in the Australian Labour Force Survey. Estimators considered include the AK -composite estimator, the early variant of regression composite estimation called $MR2$, Fuller and Rao's variant of regression composite estimation, and a BLUE estimator chosen as an "optimal" linear combination of panel estimates. An improved BLUE, obtained by calibrating the BLUE estimator to some population benchmarks, is also proposed. These estimators are compared in terms of their differences from the conventional regression estimator, their standard errors, and their usefulness for seasonal adjustment and trend estimation.

The final paper of the special section, by Gambino, Kennedy and M.P. Singh, describes the regression composite estimator that was implemented for the Canadian Labour Force Survey. This estimator is based on the work of A.C. Singh and colleagues and the compromise suggested by Fuller and Rao. The new estimators are compared to the previously used regression type estimators for a number of series. They find that the new estimators are usually more efficient and stable, and more often allow successful seasonal adjustment of the estimate series.

Kim proposes a new method for variance estimation that accounts for random imputation based on a linear regression imputation model. The method is based on creating a set of pseudo-values for y , such that a conventional variance estimator based on these pseudo-values also accounts for the imputation. Calculation of the pseudo-values is described first for simple random sampling and then for complex designs. The approach is shown to be asymptotically equivalent to the adjusted jackknife of Rao and Sitter, and properties are investigated in a simulation study.

Raghuathan, Lepkowski, Van Hoewyk and Solenberger in “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models” address the important issue of imputing into a complex data structure where explicit full multivariate models cannot be easily constructed. They adopt the approach of imputing on a variable by variable basis conditioned on all the observed variables. This implies that the imputations are created through a sequence of multiple regressions that vary depending on the type of variable being imputed.

In their article, Dufour, Gagnon, Morin, Renaud and Särndal propose a measurement of distance which can be used to measure the relative incidence of the nonresponse adjustment, calibration and the interaction between these two procedures. This measurement enables them to study and measure the change (from the initial to the final weight) resulting from the weight modification procedure. They use this measurement as a tool to compare the effectiveness of various non-response adjustment methods through a simulation study applied to the data from the Survey of Labour and Income Dynamics. The measurement is also applied to data from the National Longitudinal Survey of Children and Youth.

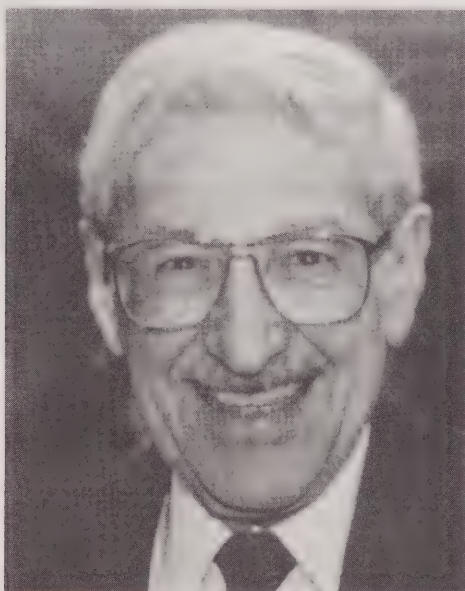
In recent years there has been an increasing number of attempts to survey homeless people in major cities. The difficulty of constructing a reliable and efficient survey frame and sampling method, and the fluidity of the population over time make surveying of this population particularly difficult. The final paper of this issue, by Ardilly and Le Blanc, describes sampling and estimation for a current survey of homelessness in France. Problems and challenges particular to this type of survey are also described. The proposed survey will sample homeless individuals indirectly by sampling the services such as shelters and meal services which they may use. The weight-share method is shown to be an effective way to obtain unbiased weights for different periods of time such as an average day or an average week.

Finally, I would like to take this opportunity to express my sincere thanks to Frank Mayda, Production Manager of *Survey Methodology*, who recently retired. His involvement with *Survey Methodology* since 1987 has been invaluable. I would also like to announce that Eric Rancourt has replaced Frank Mayda as Production Manager.

M.P. Singh

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association*. The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*.



JOSEPH WAKSBERG

Joseph Waksberg (known universally as "Joe") currently is Chair of the Board of Directors of Westat, a statistical research firm located in Rockville, MD. Throughout a career that now spans more than 60 years, he has made important contributions to sampling theory, developed innovative applications of the theory, and conducted research in a broad array of survey methodology issues. He is author or co-author of numerous papers on sampling methods, including random digit dialing, sampling for rare populations, sampling for panel and rotating design surveys, and the role of sampling in population censuses. Additional contributions have ranged from methodological research on labor force measurement, evaluation of the quality of U.S. censuses, the effects of telescoping and other problems of recall on survey results, research on the effects of cash incentives on response rates and survey costs, small area estimation, and the development of models to estimate

election night results. His goal has been to improve both survey theory and practice. Last, but not least, he has been teacher and mentor to generations of statisticians.

Born in Kielce, Poland in September 1915, Joe immigrated with his family to the United States in 1921. Shortly after graduating from the City University of New York (CUNY) in 1936 with a degree in mathematics, he moved to the Washington D.C. area and, after a brief stint with the Navy Department, joined the Census Bureau in 1940 as a clerk. He remained at the Census Bureau for 33 years, retiring in 1973 as Associate Director for Statistical Methods, Research, and Standards. In the early 1960's, Waksberg, in association with Neter, initiated a classic study on the magnitude of various types of memory recall problems. This landmark effort led to procedures for reducing the effects of recall problems through both an innovative sampling and data collection approach (Neter

and Waksberg 1964; Neter and Waksberg 1965). Joe's interest in this area has continued; for example, he helped design and analyze results from an experiment to measure the direction and magnitude of possible biases from a one year recall survey for the U.S. Fish and Wildlife Service (Chu, Eisenhower, Hay, Morganstein, Neter and Waksberg 1992). The results of that experiment had a substantial effect on the redesign of the survey. More importantly, the work also added significantly to knowledge about respondent bias when respondents are asked to recall the frequency of activities under varying recall periods, and indicated methods of minimizing the mean square errors in the design of such surveys.

The current stature of the U.S. Current Population Survey (CPS) as a model of statistical efficiency fully reflects his influence and contributions while in charge of sampling, statistical standards, and research for the Census Bureau's household survey program. Notable among the changes introduced during his tenure which bear his imprint are the improved methods of sample selection and estimation, including the use of list sampling, replication variances, determination of appropriate cluster size, treatment of rare events, and composite estimation. At the same time, he played a major role in the experimental research carried out on alternative rotation and estimation patterns, on the use of a single household respondent, and on the effects of variable recall periods on labor force measurement.

No discussion of Joe's stay at the Census Bureau is complete without some reference to his many contributions to the decennial census programs. A good example is the evaluation program for the 1970 Census, which Waksberg developed, designed, and directed. Consisting of a series of 25 separate projects, it was considered at that time as "radical"; today that program stands as the model for ongoing programs of decennial census research. When early field returns in the 1970 Census showed a serious overstatement in the reporting of "vacant" units, Waksberg designed, developed, and implemented, under great time constraints, an innovative sample survey program which revisited a sample of vacant units to estimate the proportion occupied. An adjustment procedure was then developed and applied, at the small area level, to the universe of vacant units identified in the census (Waksberg 1998). Subsequently, with the introduction of Revenue Sharing legislation in 1972, with its requirement that the Bureau produce annual estimates of population and per capita income for all 39,000 governmental units in the U.S., Waksberg proposed using administrative records in concert with survey data to provide the required local area estimates of population and per capita income. He initiated research on matching IRS records for adjacent years in order to obtain small-area (county) estimates of gross and net migration and changes in income levels, research that led to the development and implementation of a small area estimation program that is basically still in use today.

Waksberg's years at Westat, which began in 1973, first as Senior Statistician and Vice President, and recently as in-house consultant and Chair of the Board, have shown the same dedication to innovation, experimentation, and quality in meeting the needs of its clients and in developing samples and carrying out survey research. In assisting the National Center for Health Statistics in designing samples for both the National Health Interview Survey and the National Health and Nutrition Examination Survey, he made major contributions to innovative methods for efficient oversampling of minority populations, by following up work he had done earlier on this subject (Waksberg 1973). His work with Judkins and Massey provides important information on residential concentrations by race and ethnic origin, essential to assessing the usefulness of oversampling geographical areas for minority populations, and persons in poverty, another subpopulation for which oversampling is often required (Waksberg, Judkins and Massey 1997). He was a co-developer of the Mitofsky-Waksberg method of two-stage sampling of telephone households (Waksberg 1978), which became the standard approach for RDD sampling in the United States. Waksberg continued to explore ways of improving RDD sampling by examining the bias from list-assisted samples (Waksberg 1983; Brick and Waksberg 1991), which have resulted in modifications and improved efficiencies of the method and, subsequently, to a completely different method of RDD sampling (Brick, Waksberg, Kulp and Starer 1995). More recently, he participated in an examination of alternative ways of adjusting for households lacking telephones (Brick, Waksberg and Keeter 1996). His work in RDD sampling clearly demonstrates his life-long desire to constantly reexamine statistical approaches and find new methods to improve upon or even replace the standards, including those he helped establish.

Mr. Waksberg has shared his knowledge and expertise in a wide range of venues outside his office. For many years, he taught at the Graduate School of the U.S. Department of Agriculture, and was a regular lecturer at the University of Michigan summer program in sampling methods. He also has been a frequent consultant on sampling and survey techniques to governmental statistical organizations throughout the world, through the sponsorship of the U.S. Agency for International Development and the United Nations, as well as at the request of individual countries, and has provided advice to the statistical offices of China, Argentina, Brazil, Cuba, Venezuela, Turkey, and South Vietnam. He has also represented the United States at international statistical meetings, served as technical expert under UN auspices, and been a member of a team sent to South America by the American Statistical Association to coordinate activities of their national statistical societies.

He is a member of the American Statistical Association, of which he has been elected Fellow, the International Association of Survey Statisticians, and the International

Statistical Institute, and has served as a member of various panels of the National Academy of Sciences to evaluate specific Federal Statistical programs. He was the first recipient of the Roger Herriot Award, awarded by the *Washington Statistical Society* and the ASA Sections on Government Statistics and on Social Statistics for "innovation in federal statistics", and is a recipient of the Gold Medal Award of the U.S. Commerce Department. Finally, his greatest impact may be through the large number of colleagues who were inspired in their own efforts by his personal example, by his teaching, by his leadership, and by his kindness, thoughtfulness, and understanding.

REFERENCES

- BRICK, J. M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-41.
- BRICK, J. M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.
- BRICK, J. M, WAKSBERG, J., KULP, D. and STARER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.
- CHU, A., EISENHOWER, D., HAY, M. MORGANSTEIN, D., NETER, J. and WAKSBERG, J. (1992). Measuring the recall error in self-reported fishing and hunting activities. *Journal of Official Statistics*, 8, 19-39.
- NETER, J., and WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.
- NETER, J., and WAKSBERG, J. (1965). Response Errors in Collection of Expenditure Data from Household Interviews: An Experimental Study. (Bureau of the Census Technical Paper No. 11). Washington, DC: U.S. Government Printing Office.
- WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 1973, 429-434.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- WAKSBERG, J. (1983). A note on locating a special population using random digit dialing. *Public Opinion Quarterly*, 47, 576-579.
- WAKSBERG, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. (With Discussion). *Journal of Official Statistics*, 14, 119-147.
- WAKSBERG, J., JUDKINS, D. and MASSEY J. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.

2001 WAKSBERG INVITED PAPER

Author: Gad Nathan

Gad Nathan is Professor of Statistics at the Hebrew University of Jerusalem and has long been associated with the Israel Central Bureau of Statistics, most recently as Chief Scientist. He received his Ph.D. from Case Institute of Technology, Cleveland OH and has published numerous papers in leading statistical journals, including *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society*, *Survey Methodology*, *Journal of Official Statistics* and *Sankhya*. His main research areas are sampling methodology, inference from complex samples, computer assisted interviewing and telesurveys. He has held visiting and consulting positions at several academic institutions and statistical agencies in North America and in Europe and has served as Vice-President of the *International Statistical Institute* and of the *International Association of Survey Statisticians*, as well as President of the Israel Statistical Association and Chairman of the Israel Public Council of Statistics.

Telesurvey Methodologies for Household Surveys – A Review and Some Thoughts for the Future

GAD NATHAN¹

ABSTRACT

We consider 'telesurveys' as surveys in which the predominant or unique mode of collection is based on some means of electronic telecommunications – including both the telephone and other more advanced technological devices such as e-mail, Internet, videophone or fax. We review, briefly, the early history of telephone surveys, and, in more detail, recent developments in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality. All these methodological developments have led the telephone survey to become the major mode of collection in the sample survey field in the past quarter of a century. Other modes of advanced telecommunication are fast becoming important supplements and even competitors to the fixed line telephone and are already being used in various ways for sample surveys. We examine their potential for survey work and the possible impact of current and future technological developments of the communications industry on survey practice and their methodological implications.

KEY WORDS: Telephone surveys; Internet surveys; Sample design; Nonresponse; Coverage.

1. INTRODUCTION

Electronic telecommunications have become a predominant factor in practically all aspects of modern life at the beginning of the new millennium. Sample surveys are no exception and the widespread use of the telephone as a prime mode of communication for at least the past quarter of a century has had an important influence on survey practice. In fact, the telephone survey has become the major mode of collection in the sample survey field, especially in North America and Western Europe, both for surveys of households and individuals and for surveys of establishments. Other modes of advanced telecommunication, such as e-mail, Internet, videophone, fax and mobile phones are fast becoming important supplements and even competitors to the fixed line telephone. They are already being used in various ways for sample surveys and in this review paper we intend to examine their potential for survey work and the methodological implications of their use. We therefore wish to use the term 'telesurvey' for any survey in which the predominant or unique mode of collection is based on some means of electronic telecommunications – including both the telephone and other more advanced technological devices. Conventional surveys based on face-to-face interviews in the home or (snail-)mail surveys are not included, unless a substantial component of the survey is based on some telecommunications instrument. Although this paper focuses on surveys of individuals and households, much of it is relevant to establishment surveys too. We refer to telesurvey 'methodologies' in the plural, since it seems obvious that no single methodology will be suitable for use with the plethora of possible communication devices available in the future and their combinations.

This paper has been prepared in recognition of Joe Waksberg's unique contributions to survey methodology,

in general, and to telephone survey methodology in particular. It is well recognized today that his groundbreaking paper, Waksberg (1978), paved the way for the widespread efficient use of random digit dialing for telephone surveys and serves as a threshold point in the development of telesurvey methodology. Together with many of his subsequent papers, his work has had a profound influence on the theory and practice of telephone survey methodology, some of which will be examined in this paper.

We shall deal primarily with the statistical aspects of telesurvey methodology but recognize that these are not independent of non-statistical aspects, such as the cognitive features of telesurvey interviewing, survey administration and ethical considerations. In the following section we briefly review the early history of telephone surveys, through 1978. Section three reviews in some detail more recent developments in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality. Finally in section four we consider the possible impact of current and future technological developments of the communications industry on survey practice and their methodological implications.

2. THE EARLY HISTORY OF TELEPHONE SURVEYS

In the following we review briefly the overall early development of the use of telephones for survey work, as background for the developments in telesurvey methodologies to be described later. More detailed and comprehensive coverage is provided in several books and survey papers, *e.g.*, Blankenship (1977a), Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg (1988), Frey (1989),

¹ Gad Nathan, Departement of Statistics, Hebrew University, 91905 Jerusalem, Israel.

Lavrakas (1993), Casady and Lepkowski (1998, 1999) and Dillman (1978, 2000).

Telephones have been used for survey work since the thirties, though generally as a supplementary mode of collection. Some have erroneously blamed the disastrous failure of the Literary Digest survey's prediction of a landslide victory of Landon over Roosevelt in 1936, at least partially, on telephone undercoverage (Katz and Cantril 1937; Payne 1956; and Perry 1968). In fact the survey was based on mail questionnaires and although telephone lists were used as a sampling frame (in combination with lists of automobile registrations), it seems that the failure was due more to nonresponse than to frame undercoverage (Bryson 1976; Squire 1988; and Cahalan 1989).

Most of the earliest reports on the use of the telephone in survey work were in the areas of public health or in market research applications. Many of them used some combination of telephone interviewing with other modes of collection and in some cases they included empirical comparisons of response rates or outcomes in order to assess mode effects. For instance, Cunningham, Westerman and Fischhoff (1956) and Bennet (1961) report on telephone surveys for follow-up studies of patient treatment and Fry and McNaire (1958) on a national follow-up to a mail questionnaire to obtain opinions of hospital staff – all with high response rates. Mitchell and Rogers (1958) used telephone interviewing for a survey of telephone households on the consumption of dairy products and compare the results with those obtained from a control sample of non-telephone households. Cahalan (1960) compares results from telephone interviews with those from personal interviews in measuring newspaper readership with favourable results. Eastlack (1964) in a comparative telephone study of advertising recall and product usage shows that a rigorous call-back protocol provides more accurate results than a method without call-backs. Coombs and Freedman (1964) report on high telephone response (92%) in a longitudinal fertility survey, supplemented by personal interviews. Sudman (1966) describes several supplementary uses of the telephone for survey work, which include making of advance appointments and screening for rare populations, with positive results for cooperation rates and cost reductions.

In the late sixties telephone surveys really came of age, as a result of several different developments. First of all the rapid increase in telephone coverage in Western Europe and North America implied that telephone interviewing could be used as a primary mode of collection. In the US household telephone coverage reached a level of 88% in 1970 (Massey and Botman 1988) and this level was reached somewhat later in most Western European countries, in Australia and in New Zealand (Trewin and Lee 1988). In parallel to the rapid increase in telephone penetration in many countries a serious decline in response rates and difficulties in contacting respondents by door-to-door collection were experienced in the late sixties. This led to

serious consideration of telephone surveys both to reduce costs and to achieve higher cooperation rates. The use of telephone interviewing advanced most rapidly in commercial and academic survey organizations and less so in official government statistics. For instance the Federal Committee on Statistical Methodology (1984) reports that only about 11 percent of US Federal surveys in 1981 involved telephone interview in any form, in most cases in addition to other modes.

At first telephone interviewing was viewed with apprehension, even when used only as a supplementary mode of collection, due to fears of high nonresponse rates and response biases considered inherent when interviewing was not carried out face-to-face. Results of some of the earlier telephone surveys seemed to reinforce these fears. For instance, a study of leaflet receipt by Larson (1952) raises serious doubts on the validity of telephone responses on the basis of a face-to-face interview follow up. Similarly Oakes (1954) reports on suspiciously lower response on improvements to a consumer service via the telephone than obtained in face-to-face interviews. Schmiedeskamp (1962) in an attitude survey on consumer finances finds greater avoidance of taking strong positions when telephone interviewing was used. Wiseman (1972) in a comparison of mail questionnaire, telephone and face-to-face personal interviewing finds mode effects for sensitive issues (abortion and birth control). The main differences, however, are between responses to mail questionnaires and to personal interview (telephone or face-to-face).

Many of these fears were allayed at an early stage by the results of a number of more rigorous empirical studies. Thus Hochstim (1967) in a well-designed controlled experiment compares collection by mail, telephone and personal interview as the primary mode of collection. The results demonstrate convincingly that the three strategies of data collection prove to be practically interchangeable when compared with respect to rate of return, completeness of return, comparability of findings and validity of responses. The major difference between modes is with respect to cost, with a clear preference for the mail or telephone strategy. Similarly a small test carried out by Colombotos (1965) on samples of a population of physicians shows no significance differences between responses obtained by telephone and by in-person interviews. Janofsky (1971) reports similarity in willingness to express feelings on health issues between telephone respondents and face-to-face interview respondents. A well designed validation study by Locander, Sudman and Bradburn (1976) of the effects of question threat and mode of collection found no meaningful differences in response bias between telephone and face-to-face interviews. Finally, in a small carefully controlled field experiment, Rogers (1976) tested the effects of alternative interviewing strategies on the quality of responses and on field performance in a survey on a variety of complex attitudinal, knowledge and personal items. The results again indicate that the quality of data obtained by telephone is

comparable to that obtained by interviews in person. A major national study comparing telephone and face-to-face interviewing was conducted by Groves and Kahn (1979). It was based on an intensive analysis of the large omnibus surveys carried out under the two modes by the University of Michigan Survey Research Center. It provided important information on data quality which did not indicate any substantial mode effects. These and other early studies, which foreshadowed several systematic studies of mode effects carried out in the eighties and nineties (to be discussed later) contributed to the legitimacy of telephone surveys as a standard mode of collection.

The initial use of telephones for sample surveys was usually based on samples selected from general frameworks, such as telephone directories, or from specific frameworks for small sub-populations. Towards the end of the sixties there was increased awareness of high rates of unlisted telephone numbers and of substantial differences between households with listed and non-listed numbers (see details in section 3.1.1). An important development that overcame this problem was the sampling method of Random Digit Dialing (RDD), first introduced by Cooper (1964) and further improved and developed by Eastlack and Assael (1966) and by Glasser and Metzger (1972). An inherent inefficiency of these basic element RDD methods was the large amount of numbers to be called that did not yield an interview (non working and non residential numbers). A two-stage RDD sampling method was first proposed to deal with this problem by Mitofsky (1970) and subsequently elaborated and put on a firm theoretical basis by Waksberg (1978). The introduction of what was to become known as the Mitofsky-Waksberg scheme contributed greatly to the widespread use of telephone surveys in the eighties and nineties.

Finally the technological advances in telecommunications and automation in the sixties and seventies contributed to the advantages of telephone surveying. Universal direct long distance dialing enhanced the possibilities of carrying out national surveys from a single center or from a small number of interviewing centers with all the advantages of central control and administration. However the greatest impact on the expansion of telephone surveys has undoubtedly been the introduction of Computer Assisted Telephone Interviewing (CATI) in the seventies. This is due both to the simplicity of CATI for conducting telephone interviews and to the possibilities it offers for the use of automation in many important non-interviewing tasks, (*e.g.*, dialing, recall schedules *etc.*).

One of the first uses of the computer for on-line questioning was in the form of a multi-station computer-based laboratory experiment designed to elicit subjective information – Shure and Meeker (1970). A good account of the early history of CATI can be found in the special issue of *Sociological Methods and Research* (Freeman and Shanks 1983), following the Berkeley Conference on Computer-Assisted Survey Technology held in Spring 1981. Market

research organizations were the first to introduce CATI systems for their current operations. Chilton Research Services developed and used the Survey Response Processor on a current basis already in 1972 – Fink (1983). Other commercial survey organizations, applying different systems, realized early on the advantages of CATI – for instance the A&S/CATI™ system (Dutka and Frankel 1980). Academic survey research organizations were quick to follow with the earliest systems developed at UCLA and Berkeley for the large scale CATI-based California Disability Survey – Shanks, Nicholls and Freeman (1981) and Shanks (1983). Another early development of a CATI system at an academic survey organization, using a different approach, based on microcomputers, was that of the University of Wisconsin (Palit 1980; Palit and Sharp 1983). In Europe the first survey research organizations to use CATI were Social and Community Planning Research (SCPR – now the National Centre for Social Research) in the UK (Sykes and Collins 1987) and the State University of Utrecht, Netherlands (Dekker and Dorn 1984). The introduction of CATI systems into official statistics was slower. In the US it started in 1982 at the Census Bureau (Nicholls 1983) and at the National Agricultural Statistics Service (Tortora 1985) and at the same time at Statistics Netherlands (1987). By 1987 practically all organizations surveyed in a (non-probability) sample of 27 survey organizations (eighteen in the US and nine elsewhere) were using CATI for some or all of their telephone surveys – Berry and O'Rourke (1988). A report of the Federal Committee on Statistical Methodology (1990) indicated that the number of CATI installations worldwide at the end of the eighties was estimated to exceed 1,000 and that in 1988, the U.S. Government had 51 cooperating CATI centers. It should be noted that the development of CATI quickly became part of a wider movement toward computer assisted interviewing (CAI) or computer assisted information collection (CASIC), which includes also CAPI (Computer Assisted Personal Interviewing) and CASI (Computer Assisted Self Interviewing) – Nicholls (1988). A more complete history of the development of CATI and of CASIC, in general, can be found in Couper and Nicholls (1998).

3. RECENT DEVELOPMENTS IN TELEPHONE SURVEYS

In the last quarter of a century telephone surveying has definitely come of age. Lyberg and Kasprzyk (1991) claim that it has become the dominant mode of collection in countries with extensive telephone coverage.

Hundreds of scientific papers have been published during this period on a wide range of different aspects of telephone surveys. Several general books on the subject have appeared – Blankenship (1977a), Groves and Kahn (1979), Frey (1989) and Lavrakas (1993). A number of conferences have been devoted to telephone survey

methodology or have dealt with specific aspects of the topic. The results have appeared in monographs or special issues of scientific journals. A major conference on telephone survey methodology was held in November 1987 in Charlotte, NC, with the resulting volume edited by Groves, Biemer, Lyberg, Massey, Nicholls and Waksberg (1988) and the special issue of the *Journal of Official Statistics*, edited by Groves and Lyberg (1988b). The Berkeley Conference on Computer-Assisted Survey Technology held in Spring 1981 (Freeman and Shanks 1983) dealt primarily with telephone surveys and CATI was a major topic at the InterCASIC '96 International Conference on Computer Assisted Survey Information Collection, held in San Antonio, TX in December 1996 (Couper, Bethlehem, Baker, Clark, Martin, Nicholls and O'Reilly 1998) and at the ASC 3rd International Conference at Edinburgh in September 1999 (Banks, Christie, Currall, Francis, Harris, Lee, Martin, Payne and Westlake 1999).

Extensive bibliographies with several hundred entries can be found in the above sources, as well as in Khurshid and Sahai (1995), which covers the period through 1991, and in Survey Research Center (2000), which updates previous bibliographies with respect to sample design for household telephone surveys through 2000.

In the following we review the development of telephone survey methodology for household surveys during the past 25 years in the areas of sample design and estimation, coverage and nonresponse and evaluation of data quality.

3.1 Sample Design and Estimation

Sampling methodology for telephone surveys is based on the general principles of sampling. It is primarily adapted to the special situation of telephone surveys with respect to the sampling framework used. Thus we adopt the classification proposed by Lepkowski (1988) for telephone sampling methods, according to the underlying sampling framework – directory and commercial lists, telephone numbers (RDD) and combined methods (list-assisted and dual frame).

3.1.1 List-based Sampling Procedures

As mentioned above, the earliest telephone surveys were all based on samples selected from lists. In many cases they were mixed-mode surveys where telephone interviewing was used to supplement for non-response in face-to-face interviews or for follow-up. Thus so-called 'warm telephone interviewing' schemes have been used in the US Current Population Survey and in the Canadian Labour Force Survey – Drew, Choudhry and Hunter (1988). In these cases sampling is based on a general list framework to which information on telephone numbers is added and no special features of the use of the telephone are involved in the sample design. The same goes for 'pure' telephone surveys of special populations, such as physicians, for which a complete list of the population is available with telephone numbers and can be used as a sample framework

– see, for example, Gunn and Rhodes (1981). Another example is where telephone interviewing is used in follow-up waves of a panel survey with the first contact carried out by a face-to-face interview. For instance in the Israel Labor Force Survey the first contact is by a home visit and the second and third waves are carried out by telephone for households who are willing to respond by telephone – Nathan and Eliav (1988). A related approach, used recently in a pilot study for the US National Study of Health and Activity (Maffeo, Frey and Kalton 2000), is to take an area sample, find telephone numbers where possible, for telephone interviewing, and use face-to-face interviewing for other households and for telephone nonrespondents.

The most easily obtained and low-cost directory that can be used as a framework for a telephone surveys is, of course, the telephone directory itself, or some modification of it. Originally the paper version of the directory was used, while nowadays an electronic version would usually be available. The major deficiencies of the telephone directory as a sampling framework are well documented. They are undercoverage, overcoverage, duplication and lack of auxiliary information. Undercoverage is by far the most serious deficiency and includes both non-telephone households and households with telephones unlisted by choice or those not yet included in the directory. The biases due to non-telephone households are, of course, irrespective of the framework used and will be dealt with in section 3.3.

The extent of unlisted telephones varies considerably by country and type of location, as well as by other household variables. Sykes and Collins (1987) report on an unlisted rate of 4% in the Netherlands and 12% in the UK. Fréjean, Panzani and Tassi (1990) estimate the unlisted rate in France as 14% and national US estimates in the seventies were of over 17-19% (Blankenship 1977b and Glasser and Metzger 1975). Rich (1977) reports on increasing rates of nonpublished telephones (excluding those involuntarily unlisted) in the Pacific Telephone's California serving area from 9% in 1964 to 28% in 1977. In addition some 5% of home telephones in California were estimated to be involuntarily unlisted (assigned after publication of the directory). More recent studies show substantially higher unlisted rates. Thus Genesys (1996) reports unlisted rates of 40% in 1993 and of 37% in 1995, based on national samples of more than 100,000 RDD telephone interviews and Survey Sampling Inc. (1998) estimates the US national unlisted rate for 1997 at 30%. Results of a small-scale study of the Jerusalem area (Nathan and Aframian 1996) indicate an unlisted rate of 27%.

Many studies have shown substantial differences between listed and unlisted telephone household characteristics, indicating disturbing potential coverage biases for directory-based samples. In the US these differences were demonstrated, for instance, in a study by Brunner and Brunner (1971), who found highly significant differences between listed and unlisted telephone households with respect to a wide range of demographic and socio-economic

variables. Leuthold and Scheele (1971) found higher rates of nonlisting among blacks, city dwellers, young people, apartment dwellers, divorced and separated and among service workers. Similarly, Roslow and Roslow (1972) found significant differences in audience shares between listed and unlisted telephone households. Glasser and Metzger (1975) showed that nonlisted rates were higher in the West, in major metropolitan areas, among non-whites and the young. Blankenship (1977b) and Rich (1977) found highly significant differences between listed and unlisted households with respect to sex and age of household head, occupation, household size and income. In the UK Sykes and Collins (1987) found more unlisted numbers among the young, the poorest and those living in London. The results of Nathan and Aframian (1996) for the Jerusalem area showed lower rates of TV ownership and of TV viewing (of those with TV) in an RDD sample as compared with a directory listing sample.

Besides the undercoverage resulting from unlisted numbers, as indicated above, directory listings also suffer from problems of overcoverage, duplication and lack of updated auxiliary information. Overcoverage occurs when a unit outside the population is included in the framework. This may be due to the fact that disconnected numbers often remain in the directory, commercial numbers are not always clearly designated as such or other cases of unrecognized ineligibility. Duplication occurs when the same unit is represented in the frame more than once and the duplication is not recognized. Duplication can usually be discovered during sampling if the entries for the same household are listed consecutively but not if they appear separately (*e.g.*, under different surnames). If duplication is ascertained during the interview (*i.e.*, by obtaining information on the number of connected lines available to the household or the number of directory listings) it can be dealt with by appropriate weighting. Although these problems are surmountable, at a cost, that of undercoverage is not and this indicates the need for more representative sample frameworks than provided by directories. A popular alternative to the traditional telephone directory (in general prepared by the company providing telephone service to the area) has been the lists prepared by commercial firms, usually for purposes of marketing. These may be city directories, obtained from municipal address listings with telephone numbers obtained from directories or other sources, subscriber lists of telephone companies or national master address lists, such as that provided by Donelley Marketing, Inc. in the US – Lepkowski (1988). These lists provide important auxiliary data, such as geographic information, from the Census of Population and Housing and from other sources. They do not, in general, overcome the bias due to unlisted numbers and their cost may be high. They can result in some gain in sampling variance, due to the possibility of basing an efficient design on the auxiliary information. Potentially, lists used by emergency services to determine the physical location of callers could be used as

frameworks, although access to these lists would be difficult for non-government survey organizations.

3.1.2 Random Digit Dialing – The Mitofsky-Waksberg Scheme

In order to overcome many of the inherent problems of directories and commercial lists, Random Digit Dialing (RDD) methods have become a popular choice for telephone surveys, primarily in the US. These are based on the frame of all possible telephone numbers. The method was originally proposed by Cooper (1964), who added random four digit suffixes to known prefixes in a local survey. This basic element sampling method was further improved and developed by Eastlack and Assael (1966) and by Glasser and Metzger (1972), on a national level, by identifying 'working banks' of numbers from telephone company information.

The use of RDD has until recently been confined, by and large, to the US and Canada. Thus Sykes and Collins (1987) report that telephone surveys were still rare in the UK at the end of the eighties, primarily due to low telephone coverage. In particular RDD surveys were rarely used – one of the reasons being the lack of uniformity in the length of telephone numbers at the time. However recently, with the increase of telephone coverage in the UK to some 96% at the end of the nineties and the standardization of telephone numbers to ten digits, RDD surveys have become more popular – see *e.g.*, Collins (1999) and Nicolaas, Lynn and Lound (2000). Similarly, Gabler and Haeder (2000) report that an RDD method, modified in order to deal with varying telephone number lengths (from 6 to 11 digits!), is now standard procedure for telephone surveys in Germany.

Mitofsky (1970) first proposed a two-stage RDD sampling method to deal with the problem of the inherent inefficiency of these basic element RDD methods due to the large amount of numbers to be called that did not yield an interview (non working and non residential numbers). This was subsequently elaborated and put on a firm theoretical basis by Waksberg (1978) and the method became known as the Mitofsky-Waksberg scheme. This scheme or variations of it have become the predominant sampling method for telephone surveys, at least in the US.

The method is based on the fact that household telephone numbers are, in general, clustered in series of consecutive numbers or within banks of numbers with the same first r digits. For the US r is usually set at eight (for ten digit telephone numbers, including area code), so that the banks or clusters (PSU's) are of size $N = 100$ each. It is assumed that the telephone company can provide a list of all operating prefixes (area code + first three digits), *i.e.*, those to whom residential numbers have been assigned. To the six digit numbers in this list all possible choices of two digits are added, resulting in a sampling frame of eight digit numbers that represent the M PSU's in the population. Sample PSU's are selected from this frame at random (with replacement) consecutively and for each PSU selected two

final digits are selected at random. The resulting ten digit number is dialed and if the number is not that of a residence (according to the survey definition), the PSU is dropped from the sample. If it is a residence a simple random sample (without replacement) of k additional residential numbers is selected by contacting numbers selected at random (without replacement) from the PSU, until k additional residential numbers are obtained. The procedure of PSU selection continues until a fixed number of PSU's, m , has been selected. It is easily seen that, assuming that the number of residential numbers in each selected PSU, P_i , is at least k , the total sample size of residential telephone households is $m(k + 1)$ and that the final sample is an equal probability sample from the population of all residential telephone households.

Waksberg (1978) shows that if we designate by: $\pi = (\sum_{i=1}^M P_i)/(NM)$ the proportion of residential numbers in the population and by t the proportion of PSU's with no residential numbers (i.e., for which $P_i = 0$), then the expected number of total calls is given by: $m[1 + (1 - t)k]/\pi$, assuming that $P_i \geq k+1$ for all PSU's with at least one residential number. The last assumption can be dropped if PSU's are grouped so that the restriction holds in each group or if unequal weighting is used. Optimal values of the design parameters are obtained under a simple cost function and the method is extended to deal with repeated surveys. The main advantage of the method is the reduction in the expected number of calls which have to be made in order to attain a given effective sample size, especially if t , the proportion of PSU's with no residential numbers, is larger than 0.5. Groves (1977) provides data for a national study indicating a value of t of about 0.65. This advantage has to be weighed against the increase in variance due to the effect of clustering. However, taking costs into account, illustrative calculations for typical values of the parameters show that reductions in costs run between 20 and 40%.

The major operational drawback of the method is in its sequential nature. This makes it unwieldy to carry out manually. However the sequential operation poses no problem when the process of selection is fully automated. The method as described above has some additional problems, most of which can be overcome by simple modifications. Assuming that prior information on the number of telephone households is not available, selection probabilities are not known, although the value of p can be estimated from the sample. The practical necessity to introduce a stopping rule for the number of calls to numbers which do not answer or to refusals to answer, even whether the number is a residential one, implies that the method cannot be strictly applied as designed, resulting in possible bias. The problem of households with multiple telephone numbers can be overcome if correct information on the number of different lines is obtained but the required re-weighting impinges on the simplicity of equal weighting. In some cases names and addresses can be obtained for RDD

numbers by matching with address lists so that advance notice can be sent to at least part of the potential respondents. However this is a complex procedure and the difficulties in sending advance notice to respondents (common to all RDD procedures) has made the procedure difficult to consider for some official statistical agencies.

3.1.3 Modifications of the Mitofsky-Waksberg and Other RDD Methods

Some of the drawbacks of the basic method are overcome by the generalization due to Potthoff (1987a, 1987b). The method is based on the definition of a set of auspicious telephone numbers. This could consist of only residential numbers, as in the Mitofsky-Waksberg method, or a wider set which includes all residential numbers – for instance the set of all numbers which ring (including engaged, recorded messages and operators). The first stage of selection is by simple random sampling of a fixed number, m , of PSU's. From each selected PSU a fixed number of calls, c , are made and for each of them it is determined whether the number is auspicious or not. A PSU is discarded if all c numbers selected are inauspicious. Retained PSU's are defined as Type I if only one number is auspicious and as type II if two or more are auspicious. The second stage consists of selecting and dialing kc numbers from Type I PSU's and $k(c - 1)$ numbers from Type II PSU's, where k is an integer. At all dialed numbers the unit is determined as residential or out-of-scope and an interview is attempted for all residential units. A supplementary sequential segment for Type I PSU's selects additional telephone numbers that are dialed until a total of k auspicious numbers are obtained. An interview is attempted at each auspicious numbers dialed in the sequential segment. Potthoff (1987a) shows that, under certain conditions, all residential telephone numbers have the same probability of selection and develops unbiased and ratio estimates and their variances. Cost comparisons and some modifications to overcome practical problems are also given. The method reduces the problem of ambiguity on the status of dialed numbers from which no response is obtained and also the problem of exhaustion of the residential numbers in a PSU.

A large number of additional generalizations and modifications to the basic Mitofsky-Waksberg method have been proposed. Many of these attempt to reduce the burden of interviewing screening and to improve control over the initial contact sample size. Thus Hogue and Chapman (1984) propose determining cutoff points on the basis of an estimation of the probability that a PSU is 'sparse', i.e., has a small proportion of residential numbers, and propose to determine an optimal cutoff procedure on the basis of cost and variance considerations. Alexander (1988) considers two types of cutoff rules to limit interviewing screening for prefixes with low residential densities. An 'increasing rule' stops as soon as a predetermined number of calls, c_i , has been made and less than i residences have been found,

where $\{c_i\}$ is an increasing series in i . A 'decreasing rule' stops when i residences have been found if at least c_i calls have been made, where $\{c_i\}$ is a decreasing series in i . The costs for these rules are evaluated under a simple model.

Lepkowski and Groves (1986a) propose a two phase design based on matching prefixes selected in the first stage of the Mitofsky-Waksberg scheme to a commercial directory to obtain counts of listed telephones for each prefix selected. Prefixes are allocated to two strata – a low density stratum where there are no listed telephone numbers, or only a small number of them, and a high density stratum. The Mitofsky-Waksberg design is applied to the low-density stratum and telephone numbers are selected with probability proportional to the number of listed telephone numbers in the high-density stratum.

Brick and Waksberg (1991) propose using a fixed number of telephone numbers in the second stage so as to avoid sequential sampling altogether with a resulting simplicity of operation. The design, originally proposed by Waksberg (1984), is not, however, self-weighting and involves a slight bias and increased variance. Brick and Waksberg (1991) suggest considerations for the choice between the original and modified Mitofsky-Waksberg designs. For an early application of the modified Mitofsky-Waksberg method to the collection of health attitude information, apparently in an erroneous attempt to implement the original method – see Cummings (1979). Smith and Frazier (1993) compare the original and modified schemes, using data collected in the California Behavioral Risk Factor Surveillance System. The results indicate that the modified scheme speeds up the data collection, resulting in a larger sample size for the same cost. This compensates for larger design effects of the modified scheme.

Another alternative to the basic Mitofsky-Waksberg method is the use of stratification and disproportionate allocation to improve 'hit rates', proposed by Palit (1983). An evaluation of alternative treatments of unanswered telephone numbers for the Mitofsky-Waksberg design is carried out by Palit and Blair (1986). The optimal determination of parameters for the Mitofsky-Waksberg method is dealt with by Burke, Morganstein and Schwartz (1981) and the optimal allocation for the stratified version of the method by Casady and Lepkowski (1991, 1993) and by Tucker, Casady and Lepkowski (1992). Further problems relating to minimal cost allocation are treated by Palit (1983) and by Mason and Immerman (1988).

3.1.4 List-Assisted Methods

Although RDD methods overcome the undercoverage of directories due to unlisted numbers, they all still suffer from the basic problem of undercoverage due to non-telephone households (see further detail in section 3.3). In addition the lack of auxiliary information (such as geographical information), which is often available in list frames, leads to inefficiencies, even in the more sophisticated modifications

of the basic methods, mentioned above. Thus alternative methods have been sought to combine RDD samples with samples based on list and directory frames. One of the earliest attempts in this direction was that proposed by Stock (1962) and elaborated by Sudman (1973), based on replacing the last two digits of telephone numbers, selected from a directory listing, by random digits. The method was applied by Hauck and Cox (1974) to a methodological study of mode effects in screening for a special sub-population. A simpler version, popularly known as the 'Plus One' method, replaces each telephone number sampled from a directory by the number plus one (or some other digit – known as the 'plus digit method'). This supposedly overcomes the bias due to unlisted numbers. Due to its simplicity, the method has gained popularity among market researchers. However several studies – e.g., Landon and Banks (1977); and Mullet (1982) – have indicated that it is not, in fact, bias-free and also suffers from low efficiency.

Forsman and Danielsson (1997) propose a model-based approach for plus digit sampling, based on the assumption of randomly mixed listed and unlisted numbers within prefix. The model, which is tested empirically, provides model unbiased estimates. Ghosh (1984) has proposed an improved method that continues adding one to the last telephone number dialed as long as a household is not reached and stopping once a household is reached. Although still biased, the bias is reduced as compared with the simple 'plus one' method. Other list-assisted methods with RDD components, are discussed by Potter, McNeill, Williams and Waitman (1991), who stratify prefixes according to counts of published telephone numbers, while ensuring inclusion of blocks without any published numbers.

Brick, Waksberg, Kulp and Starer (1995) propose a list-assisted method that overcomes the troublesome problem of the sequential nature of the second stage sampling inherent in the Mitofsky-Waksberg scheme. The method is based on dividing the file of exchanges (100-banks) into two strata. The first consists of all exchanges with at least one listed residential phone and the second those that have none. Sampling only from the first stratum drastically reduces the proportion of nonresidential numbers which have to be dialed, but results in coverage bias. They investigate the bias and conclude that such truncated sampling methods are efficient and have operational advantages, while the resulting coverage bias (about 4%) is not too important. The method has been widely applied to replace the classical Mitofsky-Waksberg method. Similarly Statistics Canada has used the method for their General Social Survey since 1991 for the whole sample, with simple random sampling within banks of numbers identified as having at least one residential number (Norris and Paton 1991). Modifications of this design include a complete stratification of number banks on the basis of list information and using simple RDD for strata with small proportions of banks with no listing and the

Mitofsky-Waksberg method in the remaining strata. A comparison of this design with other stratified designs based on a cost model is carried out by Casady and Lepkowski (1993). Their results show that for low cost ratios (of productive selections to unproductive selections) two and three stratum RDD designs are as efficient as the Mitofsky-Waksberg scheme and that for high cost ratios they are superior.

3.1.5 Multiple Frame Designs

In an attempt to overcome some of the inherent biases of telephone surveys due to directory and telephone under-coverage, the use of dual frame mixed mode surveys, combining telephone with face-to-face interviewing, has received increasing attention. These combine conventional samples for personal interview with RDD or directory samples for telephone interviewing. Biemer (1983) investigated the optimal mix for such designs, via a simulation study, and McCarthy and Bateman (1988) propose the use of mathematical programming for attaining optimal allocation of sample units for a dual frame design, which allows posterior analysis of the effects of variations in design and cost parameters on the optimization. Choudhry (1989) proposes a cost-variable optimization for estimating proportions and Brick (1990) proposes the use of multiplicity sampling for this purpose. In a series of papers, Groves and Lepkowski (1985, 1986); Lepkowski and Groves (1984, 1986b); and Traugott, Groves and Lepkowski (1987) develop error models for these dual frame survey designs. They also report on results of experiments to compare response rates and potential biases of RDD and list samples and of several interviewing methods. The results were applied to the large scale US National Crime Survey.

Whitmore, Mason and Hartwell (1985) report on applications of dual frame dual mode methods in a US Environment Protection Agency sponsored study of personal exposure to carbon monoxide in two metropolitan areas and in a state-wide study of social service needs. In both cases commercially available directory lists were used in association with area household sampling. On the basis of an analysis of their results, they recommend the use of such dual designs in order to both benefit from the relative efficiency of telephone interviewing and to overcome the biases inherent in the use of directories as sampling frames. A combination of RDD and area sampling is reported by Waksberg, Brick, Shapiro, Flores-Cervantes and Bell (1997) for the US National Survey of America's Families in which there was particular focus on the low-income population. The nontelephone households identified in the area screening were given cellular phones for responding to telephone interviewers, thereby avoiding the need to train the area screener interviewers in a non-telephone questionnaire (Cunningham, Berlin, Meader, Molloy, Moore and Pajunen 1997).

3.2 Other Sampling Issues

3.2.1 Sampling for Special Populations

The relative low costs of telephone interviewing have made this survey mode a prime candidate for use in screening large samples in order to locate small special populations. Thus Sudman (1978) discusses the conditions under which the use of a telephone sample for screening a subgroup, to be finally interviewed face-to-face, is more efficient than face-to-face screening. By analyzing cost functions, telephone screening is found to be efficient, unless within-cluster homogeneity is small, interview densities are low and/or location and screening costs are low, relative to interview costs. Blair and Czaja (1982) propose a modification of the Mitofsky-Waksberg procedure to locate special populations that cluster geographically and describe an application to the Black population. As pointed out however by Waksberg (1983), their method requires reweighting when clusters are exhausted, which may result in reduced efficiency. This implies that the method may be efficient for the Black population but not necessarily for other minorities. Another telephone sample design targeting the US black population is proposed by Inglis, Groves and Heeringa (1987). Mohadjer (1988) proposes the stratification of prefix areas in an RDD design for sampling rare populations. The use of the Mitofsky-Waksberg method for selection of households combined with a stratified sample of individuals within household is used for the selection of a population-based control group in four epidemiological studies reported by Hartge, Brinton, Rosenthal, Cahill, Hoover and Waksberg (1984). The effectiveness of the method is studied by Perneger, Myers, Klag and Whelton (1993), on the basis of a simulation of simple random sampling, and found to be effective.

Local area surveys are another example of special populations that can be dealt with efficiently by a telephone survey. Although, in general, telephone exchanges do not define geographical areas exactly, there is a high degree of correspondence and, with some screening for those in the defined area, telephone interviewing can reduce costs considerably. For instance Banks and Hagan (1984) report on the reduction of interviewer screening by a combination of list sampling and RDD for a survey to assess the effectiveness of health programs in specific service areas. Similarly, Campbell and Palit (1988) tested a combination of list sampling and TDD – total digit dialing, using a frame of all numbers in exchanges corresponding to a given census area. They found that this resulted in a substantial saving in enumeration costs, versus face-to-face interviewing.

3.2.2 Sampling Individuals Within Households

Almost all household surveys include questions relating to individuals in the household. In some cases all individuals belonging to the household are included in the sample,

but in many cases, for various reasons, a sample of one or more individuals is selected within the household for individual questions. The classic Kish procedure (Kish 1949), predominantly used in face-to-face interview surveys raises particular problems for telephone surveys, because it requires obtaining complete household listings over the telephone. This is more difficult to obtain over the phone than in a face-to-face interview, where some of the persons may be physically present. It should be pointed out however that in many cases the information on household composition is required in any case. In addition the manipulation of the selection rules by the interviewer (*e.g.*, to accomplish high response rates), which has long been suspected in face-to-face interviewing is almost impossible in CATI surveys (where selection is invisible to the interviewer).

Toldahl and Carter (1964) proposed a method whereby only the number of persons of each sex is required. Probabilistic rules (*e.g.*, 'oldest man') are then applied to determine the individual selected, ensuring known selection probabilities for each person. However a positive probability of selection for each individual is not ensured (*e.g.*, in households with three males the one of intermediate age is never selected). The method (known as the 'Toldahl-Carter method') has been modified by Bryant (1975), in order to take into account the possibility of households with more than two individuals of the same sex. An alternative method proposed by Salmon and Nichols (1983) and by O'Rourke and Blair (1983) is to select the person with the next (or last) birthday (the 'next-birthday' or 'last-birthday' method), which ensures equal probability of selection for each household member, under the assumption that the date of interview is random. This is of course a reasonable assumption only for surveys carried out over a twelve-month period but not for surveys with shorter interview periods. This and other factors may lead to selection probabilities that are correlated with the individual characteristics. Another selection method proposed by Hagan and Meier (1983), which does not require any preliminary information on household composition, selects a pre-defined person (*e.g.*, 'eldest man'). The method again fails to ensure a positive probability of selection for each household member.

Several empirical comparisons of the above methods have been carried out. Czaja, Blair and Sebestik (1982) found no significant differences in response rates or in demographic profiles between two versions of the Toldahl-Carter method and the Kish method. Hagan and Meier (1983) compare their method, described above, with the Toldahl-Carter method and find that the method they propose has a significantly lower refusal rate, with no significant differences in demographic profiles. Salmon and Nichols (1983) compare four procedures for selecting respondents within a household unit – Toldahl-Carter, male/female alternation, next-birthday and no-selection methods – in a small telephone survey. They reach the conclusion that the next-birthday method is a relatively

efficient procedure for selecting a sample that is representative of all household members. Oldendick, Bishop, Sorenson and Tuchfarber (1988) find no significant differences between the Kish method and the last-birthday method. In a study using the last birthday method, Romuald and Haggard (1994) find that informants self-select to participate at a higher rate than expected. They investigate the effect of using memory cues on respondent self-selection and reach the conclusion that there is no significant effect. Lavrakas, Bauman and Merkle (1993) evaluate the effect of the use of the last-birthday method on within-unit coverage in a national survey and report evidence to suggest that the method leads to incorrect selection in many cases. Forsman (1993) reviews experiences of within-household sampling for 18 private opinion research companies and report on a test to compare the Kish, next/last birthday and the Toldahl-Carter methods. They conclude that the Toldahl-Carter method is somewhat better than the Kish method and that both are superior to the birthday methods. Similarly, Binson, Canchola and Catania (2000) report on a three-way comparison in a national telephone survey between the Kish, next-birthday, and last-birthday methods, and find significant differences between the three methods in the dropout rate, during the initial stages of the screening process. The Kish method had the highest dropout rates and the 'next-birthday' had the lowest rate. They conjecture that interviewers, rather than respondents, are a primary source of the higher rate of refusals when using the Kish method, due to the fact that a full household roster is required.

3.3 Coverage and Nonresponse

3.3.1 Telephone Coverage

The problem of telephone noncoverage was until very recently a major drawback of telephone surveys. Even in the US overall person undercoverage (in nontelephone households) remained at 7.2% by the end of 1986 – Thornberry and Massey (1988). By the mid-eighties household telephone undercoverage was less than 10% in most Western countries, with the highest coverage (99%) in Sweden. But some countries still had high rates of telephone undercoverage, for instance: UK 25%, Italy 29% Ireland 50%, Israel 30% – Trewin and Lee (1988). The situation changed dramatically towards the end of the century, with most Western countries reaching virtual saturation. Telephone coverage reached 94.4% in the US in 1999 (NTIA 2000); 96.6% in Australia in 1996 (St. Clair and Muir 1997); 97.0% in the UK (OFTEL 1999); 97.3% in Israel (Central Bureau of Statistics 2000); 97.9% in Finland (Kuusela and Vikki 1999); 98.2% in Canada (Statistics Canada 1999); and 99% in Germany (Federal Republic of Germany 1999).

Obviously the major problem of telephone undercoverage lies primarily in differential undercoverage rather than in its overall rate and the fact that telephone under-

coverage is highly correlated with a wide range of demographic, economic and health variables. This has been demonstrated extensively in a large number of empirical studies in the US and elsewhere – see for instance Groves and Kahn (1979), Collins (1983, 1999), Thornberry and Massey (1983, 1988), Trewin and Lee (1988) and Botman and Allen (1990). The rapid increase in overall telephone coverage over the last decade has not caused any radical change in this situation. Thus in Finland, with an overall telephone undercoverage of 2.1% in 1999, low income households (less than 675 Euros per month) had an undercoverage of 11.3% (vs. 0% for high income groups) and those living in rented accommodation 4.9% (Kuusela and Vikki 1999). In Israel telephone undercoverage was 17.9% for the lowest income decile as against 0.8% for the two highest deciles and 24.9% for single adult households with three or more children as against 2.4% for childless households with three or more adults (Central Bureau of Statistics 2000). Similarly in the US large geographical variations are still found and telephone undercoverage is found to correlate with housing deficiencies, race, education income and mobility (Shapiro, Battaglia, Hoaglin, Buckley and Massey 1996; Giesbrecht, Kulp and Starer 1996; Fox and Riley 1996; NTIA 2000). Health-related characteristics were found to differ somewhat between persons in telephone and non-telephone households in the National Health Interview Survey by Anderson, Nelson and Wilson (1998) and in the National Health and Nutrition Examination Survey by Ford (1998). However telephone coverage effects were considered to be minor in both studies.

However the main problem of telephone coverage foreseen for the near future relates to the introduction and rapid proliferation of mobile telephones. In the late nineties the proportion of households with access to at least one mobile telephone reached 76% in Finland, 59% in Denmark, 35% in Italy (Rouquette 2000) and 52% in Israel (Central Bureau of Statistics 2000). If all these mobile telephones were additional to fixed line telephones no problem would arise. However there are already strong indications of a tendency in several countries to consider the mobile telephone as an alternative to a fixed line telephone, rather than a supplement. Kuusela and Vikki (1999) report that 20% of Finnish households now have exclusively one or more mobile telephones and no fixed line and predict that within a year the number of mobile phones will exceed the number of fixed lines. Similar figures for the UK are 3% (OFTEL 2000) and for Israel 2.9% (Central Bureau of Statistics 2000). This implies that fixed line telephone coverage is down to 77% in Finland and to 94% in the UK and in Israel. In Germany it is estimated that the percentage of households with fixed line telephones will decrease to 92% by 2004 (Gabler and Haeder 2000). Furthermore the characteristics of persons with only mobile telephones are quite different from those with fixed telephone lines. In Finland, according to Kuusela and Vikki (1999), they tend

to be young, often living alone in rented apartments in urban areas. It should be noted that the transfer from fixed phone lines to mobile telephones is apparently not occurring to any large extent in North America, due to differences in pricing strategies.

Theoretically RDD sampling could be extended to mobile telephones. In practice, this may be quite difficult due to the fact that mobile telephones are by nature a personal appliance, rather than a household one. Sampling persons within a household, via a mobile telephone contact with one of the members, is well nigh impossible. Interviewing via a mobile telephone of individuals who may be anywhere is also extremely difficult. Even the determination of the total number of telephone numbers (mobile and fixed line) available to a household (required for weighting) may be daunting. We consider some possible approaches to these and other problems of the move to mobile telephones in section four.

Undercoverage of persons within covered households relates primarily to the method of selection for individuals within the household – see section 3.2.2 – and to the undercoverage due to the failure to obtain complete listings of individuals in the households. The latter effect is investigated by Makian and Waksberg (1988), by comparing data on individuals obtained from an RDD survey with those obtained from the US Current Population Survey and from the population census. They find that while mean household sizes are comparable, the RDD results are skewed towards two-person households and away from one-person households. Some of the difference could be attributed to different residence rules, but the results do not indicate undercoverage of persons in the RDD survey. They also report on an experiment in which more detailed questions were asked on household composition and found practically no improvement in accuracy of reporting. In a similar experiment, carried out by Bercini and Massey (1979), the effects of the use of names in the household roster and the position of the question on the household roster (before or after the first interview) were tested in a survey on smoking. They found that both the use of names and the position of the household roster had an effect on response and that obtaining the roster after the interview without names is optimal.

3.3.2 Nonresponse

The problem of nonresponse and the biases associated with nonresponse is basic to all survey research, but there are some specific issues of nonresponse associated with telephone interviewing. One of the major problems is the ambiguity of the results of many attempts at dialing – *e.g.*, continually engaged or no reply, numbers connected to fax machines, computer modems or answering machines. Recently automated screening devices have been developed to identify telephone numbers connected to recordings indicating whether they are not in service (Casady and Lepkowski 1999). Thus proprietary hardware and software

have been developed to detect "tri-tone" recording which indicates "not-in-service" and these numbers when dialed can be removed from the sample. Prior removal of many business phones can be carried out by matching with "Yellow Page" files. These and other methods reduce the costs of screening and the ambiguity of calls that continually receive no reply.

Technological advances, such as "call forwarding" and caller identification enhance the possibilities for non-response. In addition refusals are easier over the phone than in face-to-face interviews and breaking off the interview in its midst is also easier. These and other problems of nonresponse for 'cold' telephone interviewing and the US experience in dealing with them are reviewed extensively by Groves and Lyberg (1988a). In particular they follow CASRO (1982) and White (1983) in recommending a definition of nonresponse rates which includes in the denominator an estimate of the number of unanswered numbers that are working numbers in addition to the complete and incomplete interviews, refused eligible numbers and other noninterviewed units. The estimate of the proportion of unanswered numbers that are eligible is obtained as the proportion of answered numbers that are eligible. However this may be a biased estimator. For instance the intensive use of answering technology by businesses implies that practically all businesses will respond and can be identified as businesses. Also, as pointed out by Massey (1995), this measure has to be modified in the case of screening by defining a household screening response rate as the estimated proportion of eligible households identified as such by the screening, rather than the proportion of all households screened for eligibility. Cunningham, Brick and Meader (2000) present several detailed measures of response rates and eligibility rates for each stage of a survey with screening, as well as overall rates, in reporting on the methodology of the National Survey of America's Families.

Telephone nonresponse rates are, in general, higher than those obtained from face-to-face interviews, due to the reasons mentioned above – see Hochstim (1967), Groves and Kahn (1979), Fitti (1979), Groves and Lyberg (1988a) for US experience; Wilson, Blackshaw and Norris (1988), and Collins, Sykes, Wilson and Blackshaw (1988) for experience in UK surveys; and Drew, Choudry and Hunter (1988) for the experience of Canadian government surveys. The latter includes also comparisons of 'cold' and 'warm' telephone interviews, which show only small differences in nonresponse rates. More recently an analysis of the experience in 39 US telephone surveys carried out in the nineties (Massey, O'Connor and Krotki 1997) indicates a slight further reduction in response rates to an average of 62% and a range from 42% to 79% (though it seems that Canadian response rates have not decreased over recent years). Among the factors to which this increase in nonresponse can be attributed are the increase in the use of technological devices (answering machines, call

forwarding, multi -purpose telephone lines) and the increased prevalence of telephone solicitation, already identified as a potential problem for telephone surveys by Biel (1967). The American Statistical Association (1999) considers the effect of near saturation calling conducted by telemarketers on lowering survey cooperation rates as a serious challenge not fully addressed by survey researchers. It concludes that unless the trend can be reversed, "telephone surveys, as we know them, could disappear within the next five years". A similar view is expressed by Kalton (2000).

As is the case for telephone noncoverage, the effect of nonresponse on biases in survey estimates is made more severe by the correlation between nonresponse and many socio-economic characteristics. Groves and Lyberg (1988a) on the basis of a review of previous work identify the main correlates of telephone nonresponse. They are age (elderly persons have higher refusal rates – see also Collins *et al.* 1988) and education (higher nonresponse among lower education groups - see, *e.g.*, Cannel, Groves, Magilavy, Mathiowetz, Miller and Thornberry 1987). On the other hand, there is evidence showing that urban-rural differences in nonresponse are diminished in telephone surveys, as compared with face-to-face surveys – Groves and Kahn (1979). More recent papers on the effects of nonresponse concentrate on specific issues. Thus Diehr, Koepsell, Cheadle and Psaty (1992) investigate the relationship of response rate and other summary variables at the prefix and at the person level. They find relationships between nonresponse and age, race and family size and type. Merkle, Bauman and Lavrakas (1993) in an investigation of the impact of callbacks on the quality of survey estimates show that age and employment status are the major correlates with the number of callbacks required. Kalsbeek and Durham (1994) investigate the effect of nonresponse in a follow-up telephone survey on breastfeeding among low-income women and find that the main correlates with nonresponse are age and degree of urbanization. Finally, multilevel modeling is applied to an extensive meta-analysis of reports on inter-mode comparisons of nonresponse by Hox, DeLeeuw and Kreft (1991). The results, based on the analysis by multi-level modeling of a total of 45 studies (35 of which included a telephone component), indicate significantly lower response for telephone studies than for face-to-face studies when models with fixed slopes are used. However when random-slope models are used the difference is no longer significant.

In attempts to reduce nonresponse in telephone surveys the effect of survey operational variables on nonresponse has been investigated. Thus Sebold (1988) finds that doubling the survey period (from two to four weeks) increased the response rate by 3 percentage points in an experiment for the US National Crime Survey. Brick and Collins (1997) investigated the effect of advance letters and screening questions on response in the US National Household Education Survey. They found that a screen-out

question approach increased response rates considerably but that the advance letter did not add to the effect of screening. Other survey variables that have been found to affect response rates are interview length (Collins, *et al.* 1988) and interviewer vocal characteristics (Oksenberg and Cannel 1988). The effect of the method of selection of sample individuals on nonresponse (in particular the requirement for household rosters) has already been mentioned in section 3.2.2.

Finally, in recent years there has been a significant increase in the use of answering machines and caller ID devices for screening unwanted calls, with obvious increased potential for nonresponse. For instance, the proportion of households with answering machines in France increased from 21% in 1995 to 40% in 1999 (Rouquette 2000), the same as in Germany (Federal Republic of Germany 1999), while in the US the proportion increased from about 25% in 1988 (Tuckel and Feinberg 1991) to over 73% by 1997 (Decision Analyst 1997). However, based on a nationwide telephone survey, Tuckel and Feinberg (1991) reach the conclusion that, in comparison to other initial non-contact groups (*e.g.*, ‘no answer’ or ‘busy’), those with answering machines are more likely to respond and less likely to refuse, resulting in a contact rate which is definitely not smaller than that of other non-contacts. In fact, it seems, according to a study by Oldendick and Link (1994), that the use of answering machines to screen out survey calls is limited to some 2-3 percent. However screeners tend to be in higher income groups, urban and with higher education. Similarly, Piazza (1993) finds on the basis of extensive data from the California Disability Survey, a telephone survey with a high number of callbacks, that although answering machine owners are more difficult to contact initially, once contacted they are at least as likely to respond as those without answering machines. They point out also that reaching an answering machine ensures that a household has been reached and that its residents do not want to miss important calls. In a study by Xu, Bates and Schweitzer (1993), designed to investigate the effect of leaving messages on answering machines, households with answering machines were found to be more likely to be contacted and to complete the interview than those without answering machines. Furthermore leaving a message on the answering machine led to a significant increase in response rate and reduction in refusals. Similarly, Harlow, Crea, East, Oleson, Fraer and Cramer (1993), based on results of a controlled experiment, found that leaving a message on the answering machine led to an increase of 15% in response, after adjusting for age, interviewer and town of residence. Koepsell, McGuire, Longstreth, Nelson and van Belle (1996) carried out a randomized trial of leaving messages on answering machines and found an overall increase of 20% in response rate. Although in a similar study Tuckel and Shukers (1997) found no significant effect, the overall findings in a range of studies indicate that the increase in

the use of answering machines has a beneficial effect on survey response, probably due to their providing the possibility of leaving positive messages and thereby enabling the screening out of telemarketing calls.

Tuckel and O'Neill (1996) estimate that the percentage of US households with caller ID increased from 3% in 1992 to 10% in 1996. Based on a national study, in which the profiles of both caller ID subscribers and answering machine owners are analyzed, they reach the conclusion that these technological devices do not yet present major obstacles for telephone survey research, since their owners tend to use the screening devices primarily to screen out recognized undesirable numbers of acquaintances rather than unrecognized numbers. However, they point out that the possibility of screening will probably lead to increases in answering machine response to repeated callbacks.

3.3.3 Weighting and adjustment

Telephone surveys often require special attention to weighting and adjustment. Although sampling designs are usually based on equal probabilities of selection, in practice these are not always achieved. For instance RDD sample designs are theoretically self-weighting but in fact unequal selection probabilities may result due to the multiplicity of telephone lines (numbers) for the same household. In this case, if information is collected on the number of telephone lines to which the household is connected, the required adjustment is straightforward. Similarly reweighting is required to take into account PSU's for which the number of in-scope numbers is less than the required cluster sample size. An additional problem arises due to the fact that it is often difficult to determine whether a telephone, from which no answer can be obtained after repeated attempts, is indeed a case of in-scope nonresponse or is, in fact, out-of-scope. Other problems requiring reweighting are nonresponse, the inherent undercoverage due to non-telephone households and the obvious necessity to use some form of multiplicity estimator for multiple-frame sample designs, based on information on the frames on which the unit is represented.

These problems are dealt with for national RDD samples carried out by the US National Center for Health Statistics in a series of papers by Thornberry and Massey (1978); Botman, Massey and Shimizu (1982); and Massey and Botman (1988). They describe the weighting adjustments carried out for the RDD US National Health Interview Survey (NHIS) and for a smoking survey to account for multiple telephones per household, for telephone coverage and for nonresponse. The adjustments were based on external data for race and geographic region and on survey information on nonresponse and on multiple telephones. Several alternative adjustment and weighting procedures are compared and evaluated. Chapman and Roman (1985) compare substitution with nonresponse adjustment in a feasibility study for the RDD NHIS and report that the results with respect to bias and variance are similar. Drew

and Groves (1989) compare alternative adjustment procedures for unit nonresponse based on external administrative data, on an explicit response prediction model and on response probabilities estimated on the basis of callback data. Casady and Sirken (1980) propose a multiplicity estimator for a multiple-frame sampling design applied to data from the US National Health Interview Survey. Brick (1990) compares the multiplicity estimator with the traditional multiple frame estimator for an educational RDD survey.

Goksel, Judkins and Mosher (1991) report on adjustments, based on modeling nonresponse propensities, for a telephone follow-up of a face-to face interview in the US National Survey of Family Growth. Adjustment based on response propensities by intensity of follow-up effort and by smoking status are proposed for a Canadian survey of attitudes to smoking restrictive legislation by Bull, Pederson and Ashley (1988).

Following a comparison by Keeter (1995) of non-telephone households with 'transient' households (those who recently gained or lost telephone service), Brick, Waksberg and Keeter (1996) propose the use of data on interruptions in telephone service in order to adjust for the undercoverage due to non-telephone households. Their results indicate that such adjustment can lead to a reduction of mean square error. Hoaglin and Battaglia (1996) compare a modified poststratification method and a model-based estimation with simple poststratification for adjusting for noncoverage in an RDD survey of vaccination coverage. The modified poststratification uses national data on vaccination rates for telephone and non-telephone children in addition to demographic and socioeconomic data used for simple poststratification, while the model-based adjustment is based on a logit model to estimate the probability of residing in a telephone household. The results show gains from the use of the modified poststratification but only slight differences between the modified poststratification and the model based adjustment. A similar adjustment based on telephone interruption data is applied by Frankel, Srinath, Battaglia, Hoaglin, Wright and Smith (1999) to NHIS data and shows conclusively a substantial reduction in bias.

3.4 Data Quality – Response Errors and Mode Effects

The quality of information obtained over the telephone has always been a controversial issue. As mentioned in section 2, apprehensions on the supposed inferiority of the quality of data from telephone interviewing were allayed at an early stage, to a large degree by some of the extensive empirical appraisals carried out in the sixties and seventies. However there was still some conflicting evidence from different studies on the relative quality of telephone and face-to-face interviewing. Although the intensive analysis of large omnibus surveys carried out under the two modes by the University of Michigan Survey Research Center

(Groves and Kahn 1979), provided important information on data quality and other issues, the mode comparisons and a comparison with external data were not conclusive. In an attempt to resolve the issue, de Leeuw and van der Zoowen (1988) carried out an extensive meta-analysis of 28 major empirical studies in which comparisons of face-to-face and telephone interviewing were investigated. The studies, carried out between 1952 and 1986 on a variety of topics, were primarily from the US but some European studies were also covered. Data quality indicators used were response validity (based on validation studies), absence of social desirability bias, item response, amount of information (for open questions or check-lists) and similarity of response. The overall finding is that if there are any differences in quality between the two modes, they are definitely very minor and that other considerations, such as costs and convenience, should be used in decisions on the use of the telephone for survey work. Similar conclusions are reached for the UK by Sykes and Collins (1988), on the basis of four comparative studies; for income data in Denmark by Körmendi (1988), in a validation study, based on administrative data; and in a comparison of financial data in a Canadian Farm Financial Survey (Caron and Lavallée 1998).

Other recent studies on mode effects concentrate on specific issues and topics but reach similar conclusions. Thus Herzog and Rodgers (1988) report on a mode comparison in a study of older adults and find only small differences. Similar results are reported by Foley and Brook (1990) for a survey on the last days of life. In a study of the sensitive topic of drug use Aquilino and Lo Sciuto (1990) find almost identical results for whites, but some significant differences for blacks, even after controlling for variables possibly related to telephone undercoverage. This may be explained by results reported by Johnson, Fendrich, Shaligram and Garey (1997) for a telephone survey of drug use, which supports a social distance model of interviewer effects.

There is little doubt that interviewers have a great effect on quality, both in face to face and in telephone surveys. The use of central telephone interviewing facilities provides more opportunities to control and monitor interviewer effects than in field interviewing. Some of the issues involved are treated by Stokes and Yeh (1988), who propose a Bayesian model for interviewer effects and methods for estimating the model parameters. A beta-binomial model for the interviewer variance component and methods of estimation of its parameters are proposed by Pannekoek (1988).

An effective way of reducing response errors in face-to-face interview surveys has been the use of records provided by the respondent to verify and recall information on income, insurance, health events *etc.* Obviously, the extension of this method to telephone interviewing involves some problems, since the interviewer cannot see the documents and even asking the respondent to get them may

involve a disruptive break in the telephone interview more frequently than in a face-to-face interview. However the use of records by respondents in telephone surveys can help to reduce response bias. Battaglia, Shapiro and Zell (1996) report on an attempt to ask respondents to use vaccination records in one of the rounds of the US National Immunization Survey and to compare the information obtained with provider records. Some 47% of the respondents did in fact use vaccination records but substantial underreporting bias was still found, possibly due to the fact that the vaccination reports were not always up to date. Similar effects are found in face-to-face surveys – see Brick, Kalton, Nixon, Givens and Ezzati-Rice (2000).

4. CURRENT AND FUTURE TECHNOLOGICAL DEVELOPMENTS

Together with almost complete telephone coverage, the very intensive technological development and the diversity of communications possibilities are continuously opening up new opportunities and potentials for using novel communication options for survey work. On the other hand, some of these developments may cause difficulties for telesurveys under the conventional methodology of today. Thus the increased sophistication of filtering devices and algorithms (as a development of the simple answering machines and caller ID devices mentioned in section 3.3) may make it easier than ever for respondents not to cooperate. In the following we examine present applications and conjectured future developments and comment on the methodological problems involved in their use.

4.1 E-Mail and Web Surveys

Internet access for households has experienced a very rapid increase in recent years. For instance in the US the proportion of households with access to the Internet has risen from 26% in December 1998 to 42% in August 2000 – NTIA (2000). Other countries have reached somewhat lower levels – the UK 28% (in August, 2000 – OFTEL 2000), Canada 25%, Finland 22%, France 7% and Belgium 5% in 1999, according to Rouquette (2000), Israel 12% (in 1999 – Central Bureau of Statistics 2000) and Germany 11% (Federal Republic of Germany 1999). This rapid increase in coverage, is still far off from attaining completeness. Furthermore, there are also some indications that, together with the increase in total use, there is also a growing category of ex-users. Katz and Aspden (1998) report that the proportion of former users of the Internet increased from 8% to 11% between 1995 and 1996. However the overall increase in access has encouraged the use of e-mail and the Internet for survey work. While coverage for an e-mail survey (EMS) is comparable to that of a Web (or Internet) survey and both are based on the use of a computer self-administered questionnaire (CSAQ), there is a basic difference between these two types of

telesurveys. The e-mail survey is very similar to a mail survey, in that it is based on sending out a text questionnaire and asking the respondent to send back the completed questionnaire. The advantage over the mail survey is in cost and in the ease and simplicity of transmission and receipt. The Web survey is, in general, based on interaction between the respondent and the survey instrument, via the use of Java, XML, or a similar instrument. It allows multiple enhancements, such as colour and animation, and extensive possibilities for sophisticated skip patterns and real-time editing. The exciting potential for innovative collection systems based on ever-developing Web tools cannot yet overcome the basic problem inherent in both e-mail and Web surveys that current coverage is completely inadequate for most human populations of interest (Dillman 2000).

Nonetheless, e-mail and Internet surveys can and are being used, with varying degrees of success, for certain populations where coverage is virtually complete or in conjunction with other modes of collection. Thus Couper, Blair, and Triplett (1999) report on an experimental study comparing e-mail and regular mail for a survey of employees in several U.S. government statistical agencies. The sampled employees were randomly assigned to a mail or e-mail mode of data collection and comparable procedures were used for advance contact and follow-up of subjects across modes. The results indicated somewhat higher response rates for mail than for e-mail, but data quality (item missing data) was similar across the two modes. In field tests for the 1999 US National Study of Postsecondary Faculty both administrators and faculty were offered the choice between completing and mailing a conventional paper questionnaire or completing a CSAQ via the Web (Abraham, Steiger and Sullivan 1998). Although it may be assumed that practically all respondents had access to the Web, only 8% of responding faculty and 17% of the institution administrators opted for the CSAQ mode. The US National Science Foundation is planning to use a Web-based option in its 1999 National Survey of Recent College Graduates, under the hypothesis that most of the survey population would be relatively computer literate and have access to the Web (Meeks, Lanier, Fecso and Collins 1998). For a review of the use of CSAQ by government agencies and private survey organizations and the problems involved, see Ramos, Sedivi and Sweet (1998).

However, most current Web surveys of general populations are based on non-probability sampling – mostly by some form of self-selection. Fischbacher, Chappel, Edwards and Summerton (1999) report on a meta-analysis of 28 surveys in the health field using e-mail and the Internet. Many of these were epidemiological studies aimed at patients of specific diseases and the problem of selection bias meant that most of the results could not be generalized. One of the largest Web surveys is the WWW User Survey carried out by the Graphics Visualization and Usability

Center at Georgia Institute of Technology (Kehoe, Petkow, Sutton, Aggarwal and Rogers 1999). Although the survey population is defined as Internet users, the lack of any sample framework for this population implies that respondents had to be solicited by various methods (Web and other media announcements, advertising banners, incentive cash prizes *etc.*), rather than sampled with known probabilities. Although some 20,000 users participated, the survey report points out that the data is biased towards experienced and more frequent users and recommends the augmentation of their data with random sample surveys. In an attempt to overcome the bias inherent in basing surveys on samples of those with internet access only, some commercial survey organizations distribute devices, which let users access the Internet through television sets, to all of its panelists on an RDD sample, to ensure consistent results (Felson 2001). However Poynter (2000) predicts that by the year 2005 95% of market research surveys will be conducted via the internet but that 80% will be based on respondents who have 'opted in', rather than on probability sampling.

On the other hand, there is evidence that Web-based data collection can be applied with relative success for establishment surveys. Nusser and Thompson (1998) report on its use for the US Department of Agriculture's National Resources Inventory Surveys; Rosen, Manning and Harrel (1998) on Web-based collection from establishments for the US Current Employment Statistics Survey and Meeks *et al.* (1998) on its use for data collection from academic institutions, federal agencies and private corporations for US National Science Foundation surveys. Assuming that the problem of coverage and sampling will eventually be resolved for households and individuals, this holds hope for Web-based collection for household surveys at some point in the future.

4.2 Other Computer Self Administered Questionnaire (CSAQ) and Computer Assisted Self Interviewing (CASI) Methods

Couper and Nichols (1998) differentiate between computer self administered questionnaire (CSAQ) collection, in which an interviewer is not present, and computer assisted self interviewing (CASI), in which an interviewer is present or delivers the survey instrument. Thus both e-mail and Internet surveys are based on CSAQ with the assistance of telecommunications technology. Other CSAQ methods are touchtone data entry (TDE), whereby respondents enter data using their touchtone telephones, and interactive voice recognition (IVR) or voice recognition entry (VRE). Both are based on respondents initiating calls to report at their convenience, after initial contact has been established, and have been extensively tested and successfully used by the US Bureau of Labor Statistics for data collection from establishments for its Current Employment Statistics program – Werking, Tupek and Clayton (1988), Winter and Clayton (1990) and Clayton

and Winter (1992). Phipps and Tupek (1991) report on a study of the quality of TDE collection, by means of a record check. Their results show that there are few problems with the method and that response errors diminish with experience. More recently US statistical agencies have initiated tests of the possibility of applying these CSAQ methods to household surveys. McKay, Robison and Malik (1994) report on initial laboratory testing of TDE for the Current Population Survey. Malakhoff and Appel (1997) report on the development of an IVR prototype at the US Bureau of Census, albeit for a listing operation by field staff. It should be noted that while TDE is obviously unique to telephone surveys, IVR could be used for other modes of collection.

Computer assisted self interviewing (CASI) methods include audio (ACASI) and video (VCASI) modes of collection and have long been regarded as the natural extensions of mail surveys that benefit from modern day technology (Dillman 2000). Their usefulness has been especially emphasized for surveys of sensitive and embarrassing topics, where the presence of the interviewer during the interview may make respondents reluctant to answer in a face-to-face interview. For a review of recent advances in these methods see Baker (1998), O'Reilly, Hubbard, Lessler, Biemer and Turner (1994), Rogers, Miller, Forsyth, Smith and Turner (1996) and Tourangeau and Smith (1998). Practically all the reported applications are of surveys in which the survey instrument is brought to the respondent's home by field staff. The use of the telephone for ACASI (T-ACASI) collection has already been tried – Turner, Forsyth, O'Reilly, Cooley, Smith, Rogers and Miller (1998). The long-expected development of videotelephony to become a widespread common form of telephone service for households has not yet materialized. If and when it occurs it should make telephone VCASI (T-VCASI) possible in the future, with important implications for telesurvey work. The addition of a visual element will help to overcome many of the problems of present day telephone surveys that are not present in face-to-face interviews (eye contact with the interviewer, use of cue cards and other visual aids). The use of videotelephony will probably not be universal for a very long time, so that at least for the time being, T-VCASI will only be able to serve as a supplementary mode of collection.

4.3 Mobile Telephones

The problems envisaged for coverage of fixed line RDD surveys due to the rapid proliferation of mobile telephones have been mentioned in section 3.3.1. In the future it is obvious that mobile telephones will have to be used to reach the ever-increasing numbers of households without fixed telephone lines. Present levels of mobile telephone coverage imply that mobile telephone surveys can, in general, only be used for specific populations or for supplementing fixed line RDD surveys. For instance Perone, Matrundola and Soverini (1999) report on a mobile telephone survey for a naturally accessible population - that

of mobile telephone subscribers in order to assess customer satisfaction. Refusal rates were found not to exceed those found in fixed line telephone surveys. However, non-contact rates were high, primarily due to subscribers being outside the signal range or shutting down their telephones. An additional problem associated with mobile phone surveys is that in many cases in North America the subscriber has to pay for received calls – Casady and Lepkowski (1999).

As mentioned above, Cunningham, *et al.* (1997) report on the use of mobile telephones to interview nontelephone households (primarily in rural areas), with the mobile telephone brought to the respondent by field interviewers. This was designed to minimize mode effects by having telephone interviews conducted by the same interviewers as those conducted for telephone households. The response rates were high, even though in some cases the interviews had to be conducted outdoors in order to obtain reasonable reception. The most intensive use of mobile phones for household surveys is no doubt for the Finish Labour Force Survey – Kuusela and Notkola (1999). Out of some 97% of interviews completed by telephone, over 20% are carried out by mobile telephone. Although the average duration of mobile telephone interviews is somewhat longer than those of conventional telephone interviews, this is probably due to socio-demographic differences between the respondent groups.

4.4 Future Technological Developments and their Effect on Telesurvey Methodology

The rapid advances in technological developments in the areas of telecommunications and information systems make it very difficult to forecast their influence on survey work. Not all these technological changes will necessarily increase the potential for using advanced telecommunications technology for survey work. The problems raised by persons who have opted to 'drop-out' from the Internet (Katz and Aspden 1998) or from fixed line telephone service (see *e.g.*, Gabler and Haeder 2000; and Kuusela and Vikki 1999) have already been mentioned. Furthermore, in some areas, such as market research and official statistics, technological developments may lead to a reduced reliance on surveys to gather information for decision-making. Thus Baker (1998) and Poynter (2000) predict that techniques such as data mining of existing data resources may become predominant for market research. Similarly, Scheuren and Petska (1993) discuss the possibilities for the use of administrative record systems for official statistics. However, there still remain important areas (for instance for opinions and unobservable behaviour) in which surveys will remain the predominant source of data. The technological advances will open new possibilities for telesurvey work, though the required methodology might become more complex than that used today.

One of the expected developments forecast for the near future is the integration of multiple communication devices

and methods – telephony (fixed line and wireless), fax, internet, e-mail, videotelephony, data transmission, television transmissions *etc.* – Baker (1998). This implies that each individual will have access to a variety of telecommunication services possibly via the same physical instrument, which could be a mobile phone (*e.g.*, via WAP technology), a PC or a TV set or some combination of these. Similarly, the survey taker may be able to gain access to respondents via several different modes. See Ranta-aho and Leppinen (1997) for some of the issues involved in this plethora of possible avenues of access. It is envisaged that the recipient will have a large degree of control over whether to receive communications at all and, if, so by which mode. This is already now ensured for many users by means of sophisticated devices for screening, forwarding, message transfer, multiple message transmission *etc.* On the other hand, the degree of control of mode of transmission by the sender will probably decrease as a result.

The implications of these developments for survey work are that mixed mode surveys and possibly multi-frame methodology will have to become predominant. Although we consider that overall telecommunications coverage will increase to some saturation point that is close to universal coverage, it seems unlikely that any given mode of telecommunication will by itself provide virtual complete coverage. Furthermore, even when a single mode may provide practically complete coverage, it is not clear that a mixed mode approach, taking into account respondents' mode preferences, is not preferable. The increased reliance of survey work on the voluntary cooperation of respondents practically dictates that we should offer the respondent the choice of mode. However it should be pointed out that mixed mode surveys are very expensive and that the present technology does not allow the simple transfer of questionnaires developed for one mode (*e.g.*, the CAI Blaise questionnaire) to another mode – *e.g.*, to a paper form.

The major problem that the new developments in telecommunications pose for survey design will probably be the choice of relevant frameworks and the allocation of sample units to modes of collection. Eventually it is envisaged that each individual will have a unique, permanent, personal communication number (or ID) through which he/she can be reached by a multiplicity of modes (written, oral or visual), via a variety of fixed line or wireless devices which could be at home, in the office or mobile. The choice of mode will be ultimately controlled by the joint decision of recipient and sender. While the idea of such a universal number (which would basically be an identity number) is no doubt anathema to libertarians, there is little doubt that it will eventually become acceptable, even if small activist groups may attempt to evade its use and even disrupt its proliferation. In fact standard universal identity number systems have been operating and are well accepted for several decades in many countries in Northern Europe and in Israel. The identity number in these countries is not regarded as confidential information and is widely used for

many administrative and commercial purposes. For example, in Israel personal cheques are required by law to include the person's ID number, name, address and telephone number.

Once such a system of unique communication numbers is operable, standard methods of sampling can be used. It may well be that complete lists of these numbers will be generally available – possibly with only limited geographical or other information. This is the situation with respect to ID's in many national registration systems. There are reasons to expect that a similar situation may prevail for communication numbers – initially at least in Europe rather than in North America. This could come about since the need for unlisted status might well be made redundant because of sophisticated screening techniques. Although screening may enhance the ease of non-response, the possibility of transmitting prior written messages by e-mail or voice mail could reduce the problem.

Sampling from such lists would be simple but in most cases might be inefficient, since it could benefit only marginally from auxiliary information. While differentiation between personal and business contacts might be ensured by the listings, it is doubtful that any household information would be available. This dictates that the sampling and reporting unit would be the individual rather than a household. This is in any case the aim of many surveys and the usefulness of the household as a sampling unit for telesurveys is definitely doubtful, even under current practice. Household information, if required, would have to be obtained from the individual and include information on household size to ensure proper weighting for household characteristics. If the communications numbering system ensures the allocation of a single number to each individual, no information is required on the modes of communication or their multiplicity.

If listings of communication numbers are not available or if the problem of unlisted numbers does persist, some form of RDD will have to be used. This should not differ much from the RDD techniques currently employed. Assuming that the communication numbering system is indeed unique and universal and also arranged by some logic, efficient methods for sampling could easily be developed. Hopefully the numbering system will still bear some relationship to geography, via the individual's permanent address. Otherwise local or even national RDD surveys will become extremely difficult to design efficiently. If sufficient information on the numbering system is available, the extent of out-of-scope numbers could be minimized.

Since it is likely that choice of the mode of communications will be largely under the control of the recipient, the question of allocation of sample units to mode of communication will probably hardly arise. The survey taker will have to prepare a whole range of collection instruments suitable for the different modes of communication. These would have to include written instruments, such as faxed, e-mail

and Internet versions of questionnaires, oral instruments, such as traditional voice interviews and automated interviewing, and combinations of these. The integration of the data obtained from these modes of collection into a uniform data set would be a formidable but surmountable technological challenge.

The almost utopian situation described above will probably take a long time to reach and in the interim suitable methodologies will have to be developed to deal with the problems arising from the short-term developments in communications technology and their application. The necessity to move from telephone surveys based uniquely on fixed line telephones to some combination of mobile and fixed-line telephone situation will have to be dealt with very shortly, as pointed out in section 4.3. Basically multiple frame methodology developed to cover both telephone households and non-telephone households can easily be extended to deal with this. The development of suitable frames and/or RDD sampling methods for mobile telephones still has to be carried out, but the necessary principles are available. The problem of combining data obtained from mobile phones which are basically personal devices with that obtained from fixed-line telephones, which are still fundamentally household devices, will have to be worked out to ensure proper weighting. To ensure this, sufficiently complete information on all the communication devices available to the household is required.

In conclusion, the advances in telesurvey methodology over the past few decades, which have made telephone surveys a viable and predominant survey instrument, will have to be continually updated to deal with the ever-changing developments in telecommunications technology and its usage. However the basic elements for these new developments are available and will continue to allow the use of advanced options to obtain high quality survey data

REFERENCES

- ABRAHAM, S.Y., STEIGER, D.M. and SULLIVAN, C. (1998). Electronic and mail self-administered questionnaires: A comparative assessment of use among elite populations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 833-841.
- ALEXANDER, C.H. (1988). Cutoff rules for secondary calling in a random digit dialing survey. *Telephone Survey Methodology*, (R.M. Groves, *et al.* - Eds.). New York: John Wiley and Sons, 113-126.
- AMERICAN STATISTICAL ASSOCIATION (1999). *More About Telephone Surveys*. ASA series: what is a survey? Section on Survey Research Methods [<http://www.amstat.org/sections/srms/brochures/telephone.pdf>].
- ANDERSON, J.E., NELSON, D.E. and WILSON, R.W. (1998). Telephone coverage and measurement of health risk indicators: data from the National Health Interview Survey. *American Journal of Public Health*, 88, 1392-1395.

- AQUILINO, W.S., and LO SCIUTO, L.A. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly*, 54, 362-395.
- BAKER, R.P. (1998). The CASIC future. *Computer Assisted Survey Information Collection*. (M.P. Couper, et al. - Eds.). New York: John Wiley and Sons, 583-604.
- BANKS, M.J., and HAGAN, D.E. (1984). Reducing interviewer screening and controlling sample size in a local-area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 271-273.
- BANKS, R., CHRISTIE, C., CURRALL, J., FRANCIS, J., HARRIS, P., LEE, B., MARTIN, J., PAYNE, C. and WESTLAKE, A. (Eds.) (1999). ASC'99 - Leading Survey & Statistical Computing into the New Millennium. *Proceedings of the ASC International Conference*. Association for Survey Computing Chesham, Bucks, UK.
- BATTAGLIA, M.P., SHAPIRO, G. and ZELL, E.R. (1996). Substantial response bias may remain when records are used in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 452-455.
- BENNET, C.T. (1961). A telephone interview: A method for conducting a follow-up study. *Mental Hygiene*, 45, 216-220.
- BERCINI, D.H., and MASSEY, J.T. (1979). Obtaining the household roster in a telephone survey: The impact of names and placement on response rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 136-140.
- BERRY, S.H., and O'ROURKE, D. (1998). Administrative designs for centralized telephone survey centers: Implications of the transition to CATI. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 457-474.
- BIEL, A. L. (1967). Abuses of survey research techniques: the phony interview. *Public Opinion Quarterly*, 31, 298.
- BIEMER, P.P. (1983). Optimal dual frame sample design: Results of a simulation study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 630-635.
- BINSON, D., CANCHOLA, J.A. and CATANIA, J.A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics*, 16, 53-59.
- BLAIR, J., and CZAJA, R. (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.
- BLANKENSHIP, A.B. (1977a). *Professional Telephone Surveys*. New York: McGraw Hill.
- BLANKENSHIP, A.B. (1977b). Listed versus unlisted numbers in telephone-survey samples. *Journal of Advertising Research*, 39-42.
- BOTMAN, S.L., and ALLEN, K. (1990). Some effects of undercoverage in a telephone survey of teenagers. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-400.
- BOTMAN, S.L., MASSEY, J.T. and SHIMIZU, I.M. (1982). Effect of weighting adjustments on estimates from a random-digit-dialed telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 139-144.
- BRICK, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methodology, American Statistical Association*, 296-301.
- BRICK, J.M., and COLLINS, M.A. (1997). A response rate experiment for RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1052-1057.
- BRICK, J.M., KALTON, G., NIXON, M., GIVENS, J. and EZZATI-RICE, T. (2000). Statistical issues in a record check study of childhood immunizations. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference: Statistical Policy Working Paper*, 30, 625-634.
- BRICK, J.M., and WAKSBERG, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, 17, 27-42.
- BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.
- BRICK, J. M., WAKSBERG, J., KULP, D. and STARER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.
- BRUNNER, J.A., and BRUNNER, G.A. (1971). Are voluntary unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.
- BRYANT, B.E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, 12, 129-135.
- BRYSON, M.C. (1976). The literary digest poll: Making of a statistical myth. *The American Statistician*, 30, 184-185.
- BULL, S.B., PEDERSON, L.L. and ASHLEY, M.J. (1988). Intensity of follow up; effects on estimates in a population telephone survey with an extension of Kish's (1965) approach. *American Journal of Epidemiology*, 127, 552-561.
- BURKE, J., MORGANSTEIN, D. and SCHWARTZ, S. (1981). Toward the design of an optimal telephone sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 448-453.
- CAHALAN, D. (1960). Measuring newspaper readership by telephone: two comparisons with face to face interviews. *Journal of Advertising Research*, 1, 2, 1-6.
- CAHALAN, D. (1989). Comment: The digest poll rides again! *Public Opinion Quarterly*, 53, 129-133.
- CAMPBELL, J., and PALIT, C.D. (1988). Total digit dialing for a small area census by phone. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 549-551.
- CANNELL, C.F., GROVES, R.M., MAGILAVY, L.J., MATHIOWETZ, N.A., MILLER, P.V. and THORNBERRY, O.T. (1987). An experimental comparison of telephone and personal health interview surveys. *Vital and Health Statistics, Series 2*, 106, Public Health service.
- CARON, P., and LAVALLÉE, P. (1998). Comparison study on the quality of financial data collected through personal and telephone interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 208-213.

- CASADY, R.J., and LEPKOWSKI, J.M. (1991). Optimal allocation for stratified telephone survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 111-116.
- CASADY, R.J., and LEPKOWSKI, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.
- CASADY, R.J., and LEPKOWSKI, J.M. (1998). Telephone Sampling. *Encyclopedia of Biostatistics*. New York: John Wiley and Sons, 4498-4511.
- CASADY, R.J., and LEPKOWSKI, J.M. (1999). Telephone Sampling. *Sampling of Populations: Methods and Applications - third edition*, (P.S. Levy and S. Lemeshow, Eds.). New York: John Wiley and Sons, 455-479.
- CASADY, R.J., and SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-609.
- CASRO (1982). *Report of the Council of American Survey Research Organization Completion Rate Task Force*. New York: Audits and Surveys Inc. (Unpublished report).
- CENTRAL BUREAU OF STATISTICS (2000). *The Household Expenditure Survey 1999*. Special Publication 1147. Jerusalem.
- CHAPMAN, D.W., and ROMAN, A.M. (1985). An investigation of substitution for an RDD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 269-274.
- CHOUDHRY, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.
- CLAYTON, R.L., and WINTER, D. L. S. (1992). Speech data entry: results of a test of voice recognition for survey data collection. *Journal of Official Statistics*, 8, 377-388.
- COLLINS, M. (1983). Computer assisted telephone interviewing in the UK. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 636-639.
- COLLINS, M. (1999). Editorial: sampling for UK telephone surveys. *Journal of the Royal Statistical Society, A* 162, 1-4.
- COLLINS, M., SYKES, W., WILSON, P. and BLACKSHAW, N. (1988). Nonresponse: the UK experience. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 213-232.
- COLOMBOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.
- COOMBS, L., and FREEDMAN, R. (1964). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.
- COOPER, S. L. (1964). Random sampling by telephone: an improved method. *Journal of Marketing Research*, 1, 45-48.
- COUPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z.F., MARTIN, J., NICHOLLS, W.L., II and O'REILLY, J.M. - (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley and Sons.
- COUPER, M.P., BLAIR, J. and TRIPLETT, T. (1999). A comparison of mail and e-mail for a survey of employees in US statistical agencies. *Journal of Official Statistics*, 15, 39-56.
- COUPER, M.P., and NICHOLLS, W.L., II (1998). The history and development of computer assisted survey information collection methods. *Computer Assisted Survey Information Collection*, (M.P. Couper, et al. - Eds.). New York: John Wiley and Sons, 1-22.
- CUNNINGHAM, J.M., WESTERMAN, H.H. and FISCHOFF, J. (1956). A follow-up study of patients seen in a psychiatric clinic for children. *American Journal of Orthopsychiatry*, 26, 602-610.
- CUNNINGHAM, P., BERLIN, M., MEADER, J., MOLLOY, K., MOORE, D. and PAJUNEN, S. (1997). Using cellular telephones to interview nontelephone households. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 250-254.
- CUNNINGHAM, P., BRICK, J.M. and MEADER, J. (2000). 1999 *NSAF In-Person Survey Methods Report No. 5*. Washington, DC: Urban Institute. [http://newfederalism.urban.org/nsaf/methodology_rpts/1999_Methodology_5.pdf].
- CUMMINGS, M.K. (1979). Random digit dialing: a sampling technique for telephone surveys. *Public Opinion Quarterly*, 43, 233-244.
- CZAJA, R., BLAIR, J., and SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques. *Journal of Marketing Research*, 19, 381-385.
- DECISION ANALYST (1997). More households using answering machines. *News Release, October 15, 1997*. [http://www.decisionanalyst.com/publ_data/1997/ansmachi.htm].
- DEKKER, F., and DORN, P.K. (1984). Computer Assisted Telephone Interviewing: A research project in the Netherlands. Paper presented at: *Conference of the Institute of British Geographers*.
- DE LEEUW, E.D., and VAN DER ZOOWEN, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 283-299.
- DIEHR, P., KOEPESELL, T.D., CHEADLE, A., and PSATY, B.M. (1992). Assessing response bias in random-digit dialing surveys: The telephone-prefix method. *Statistics in Medicine*, 11, 1009-1021.
- DILLMAN, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley and Sons.
- DILLMAN D. A. (2000). *Mail and Internet Surveys: The Total Design Method* (2nd edition). New York: John Wiley and Sons.
- DREW, J.D., CHOUDHRY, G.H. and HUNTER, L.A. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 233-246.
- DREW, J.H., and GROVES, R.M. (1989). Adjusting for nonresponse in a telephone subscriber survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 452-456.
- DUTKA, S., and FRANKEL, L. R. (1980). Sequential survey design through the use of computer assisted telephone interviewing. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 73-76.

- EASTLACK, J.O., JR. (1964). Recall of advertising by two telephone samples. *Journal of Advertising Research*, 4, 25-29.
- EASTLACK, J.O., JR., and ASSAEL, H. (1966). Better telephone surveys through centralized interviewing. *Journal of Advertising Research*, 6, 1, 2-7.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1984). *The Role of Telephone Data Collection in Federal Statistics*. Statistical Policy Working Paper 12, Washington, D.C.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1990). *Computer Assisted Survey Information Collection*, Statistical Policy Working Paper 19, Washington, D.C.
- FEDERAL REPUBLIC OF GERMANY (1999). Continuous family budget surveys for January 1999. Statistisches Bundesamt: *Press release*, 20 December, 1999. [<http://www.statistik-bund.de/presse/englisch/pm1999/p4350024.htm>].
- FELSON, L. (2001). Netting limitations: online researchers' new tactics for tough audiences. *Marketing News (American Marketing Association)*, 35, 5 [<http://www.ama.org/pubs/article.asp?id=4881>].
- FINK, J.C. (1983). CATI's first decade: The Chilton experience. *Sociological Methods and Research*, 12, 153-168.
- FISCHBACHER, C., CHAPPEL, D., EDWARDS, R. and SUMMERTON, N. (1999). The use and abuse of the Internet for survey research. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 501-507.
- FITTI, J.E. (1979). Some results from the telephone health interview system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 244-249.
- FOLEY, D.J., and BROCK, D.B. (1990). Comparison of in-person and telephone responses in a survey of the last days of life. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 382-386.
- FORD, E.S. (1998). Characteristics of survey participants with and without a telephone; findings from the third National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 51, 55-60.
- FORSMAN, G. (1993). Sampling individuals within households in telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1113-1118.
- FORSMAN, G., and DANIELSSON, S. (1997). Can plus digit sampling generate a probability sample? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 958-963.
- FOX, A., and RILEY, J. P. (1996). Telephone coverage, housing quality and rents: RDD survey biases. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 515-519.
- FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. and SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 934-937.
- FREEMAN, H.E., and SHANKS, J. M. - Eds. (1983). The emergence of computer-assisted survey research. *Sociological Methods and Research*, 12 (special issue), 115-230.
- FRÉJEAN, M., PANZANI, J-P. and TASSI, P. (1990). Les ménages inscrits en liste rouge et les enquêtes par téléphone. *Journal de la Société de Statistique de Paris*, 131, Nos. 3-4, 86-102.
- FREY J. H. (1989). *Survey Research by Telephone* (2nd edition). Beverly Hills, CA: Sage Publications.
- FRY, H.G., and MCNAIRE, S. (1958). Data gathering by long distance telephone. *Public Health Records*, 73, 831-835.
- GABLER, S., and HAEDER, S. (2000). Telephone sampling in Germany. *Paper presented at Fifth International Conference of Social Science Methodology*, Köln.
- GENESYS (1996). Unlisted numbers: what's really important. *Genesys News (Genesys Sampling Systems, Fort Washington, PA)*, 1-2.
- GHOSH, D. (1984). Improving the plus 1 method of random digit dialing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 285-288.
- GIESBRECHT, L.H., KULP, D.W. and STARER, A.W. (1996). Estimating coverage bias in RDD samples with current population survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 503-508.
- GLASSER, G.J., and METZGER, G.D. (1972). Random digit dialing as a method of telephone sampling. *Journal of Marketing Research*, 9, 59-64.
- GLASSER, G. J., and METZGER, G. D. (1975). National estimates of nonlisted telephone households. *Journal of Marketing Research*, 12, 359-361.
- GOKSEL, H., JUDKINS, D.R. and MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 581-586.
- GROVES, R.M. (1977). *An Empirical Comparison of Two Telephone Sample Designs*. Unpublished report of the Survey Research Center, the University of Michigan, Ann Arbor, MI.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II and WAKSBERG, J. - Eds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- GROVES R.M., and KAHN, R.L. (1979). *Surveys by Telephone: A National Comparison With Personal Interview*. New York: Academic Press.
- GROVES, R.M., and LEPKOWSKI, J.M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics*, 1, 264-286.
- GROVES, R.M., and LEPKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 340-345.
- GROVES, R.M., and LYBERG, L.E. (1988a). An overview of nonresponse issues in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 191-211.
- GROVES, R.M., and LYBERG, L.E. - Eds. (1988b). Telephone survey methodology. *Journal of Official Statistics* (special issue), 4, 283-416.
- GUNN, W.J., and RHODES, I.N. (1981). Physician response rates to a telephone survey: effects of monetary incentive level. *Public Opinion Quarterly*, 45, 109-115.

- HAGAN, D. E., and MEIER C. C. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly*, 47, 547-556.
- HARLOW, B.L., CREA, E.C., EAST, M.A., OLESON, B., FRAER, C.J. and CRAMER, D.W. (1993). Telephone answering machines: the influence of leaving messages on telephone interviewing response rates. *Journal of Epidemiology*, 4, 380-383.
- HARTGE, P., BRINTON, L.A., ROSENTHAL, J.F., CAHILL, J.I., HOOVER, R.N. and WAKSBERG, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.
- HAUCK, M., and COX, M. (1974). Locating a sample by random digit dialing. *Public Opinion Quarterly*, 38, 253-260.
- HERZOG, A.R., and RODGERS, W.L. (1988). Interviewing older adults: mode comparison using data from a face-to-face survey and a telephone resurvey. *Public Opinion Quarterly*, 52, 84-99.
- HOAGLIN, D.C., and BATAGLIA, M.P. (1996). A comparison of two methods of adjusting for noncoverage of nontelephone households in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 497-502.
- HOCHSTIM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.
- HOGUE, C.R., and CHAPMAN, D.W. (1984). An investigation of PSU cutoff points for a random digit dialing survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 286-291.
- HOX, J.J., DE LEEUW, E.D. and KREFT, I.G.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. *Measurement errors in surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz and S. Sudman - Eds.). New York: John Wiley and Sons, 439-461.
- INGLIS, K.M., GROVES, R.M. and HEERINGA, S.G. (1987). Telephone sample designs for the U.S. Black household population. *Survey Methodology*, 13, 1-14.
- JANOFSKY, A.I. (1971). Affective self-disclosure in telephone versus face-to-face interviews. *Journal of Humanistic Psychology*, 11, 93-103.
- JOHNSON, T., FENDRICH, M., SHALIGRAM, C. and GAREY, A. (1997). A comparison of interviewer effects models in an RDD telephone survey of drug use. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 887-892.
- KALSBECK, W.D., and DURHAM, T.A. (1994). Nonresponse and its effects in a followup telephone survey of low-income women. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 943-948.
- KALTON, G. (2000). Developments in survey research in the past 25 years. *Survey Methodology*, 26, 3-10.
- KATZ, D., and CANTRIL, H. (1937). Public opinion polls. *Sociometry*, 1, 155-179.
- KATZ, J.E., and ASPDEN, P. (1998). Internet dropouts in the USA. *Telecommunications Policy*, 22, 4/5, 327-339.
- KEETER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.
- KEHOE, C., PITKOW, J., SUTTON, K., AGGARWAL, G. and ROGERS, J.D. (1999). *Results of Gvu's Tenth World Wide Web User Survey*. Atlanta, GA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology. [http://www.gvu.gatech.edu/user_surveys].
- KHURSHID, A., and SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KOEPSSELL, T.D., MCGUIRE, V., LONGSTRETH, Jr., W.T., NELSON, L.M. and VAN BELLE, G. (1996). Randomized trial of leaving messages on telephone answering machines for control recruitment in an epidemiological study. *American Journal of Epidemiology*, 144, 704-706.
- KÖRMENDI, E. (1988). The quality of income information in telephone and face to face surveys. *Telephone Survey Methodology* (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 341-356.
- KUUSELA, V., and NOTKOLA, V. (1999). Survey quality and mobile phones. Paper presented at *International Conference on Survey Nonresponse*, Portland OR.
- KUUSELA, V., and VIKKI, K. (1999). Change of telephone coverage due to mobile phones. Paper presented at *International Conference on Survey Nonresponse*, Portland OR.
- LANDON, E.L., JR., and BANKS, S.K. (1977). Relative efficiency and bias of plus-one telephone sampling. *Journal of Marketing Research*, 14, 294-299.
- LARSON, O. N. (1952). The comparative validity of telephone and face-to-face interviews in the measurement of message diffusion from leaflets. *American Sociological Review*, 17, 471-476.
- LAVRAKAS, P.J. (1993). *Telephone Survey Methods: Sampling, Selection and Supervision* (2nd edition). Newbury Park, CA: Sage Publications.
- LAVRAKAS, P.J., BAUMAN, S.L. and MERKLE, D.M. (1993). The last-birthday method and within-unit coverage problems. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1107-1112.
- LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 73-98.
- LEPKOWSKI, J.M., and GROVES, R.M. (1984). The impact of bias on dual frame survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 265-270.
- LEPKOWSKI, J.M., and GROVES, R.M. (1986a). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 357-362.
- LEPKOWSKI, J.M., and GROVES, R.M. (1986b). A mean square error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.

- LEUTHOLD, D.A., and SCHEELE, R. (1971). Patterns of bias in samples based on telephone directories. *Public Opinion Quarterly*, 35, 249-257.
- LOCANDER, W., SUDMAN, S. and BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.
- LYBERG, L., and KASPRZK, D. (1991). Data collection methods and measurement error: an overview. *Measurement Errors in Surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz and S. Sudman - Eds.). New York: John Wiley and Sons, 237-258.
- MCCARTHY, W.F., and BATEMAN, D.V. (1988). The use of mathematical programming for designing dual frame surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 652-653.
- MCKAY, R.B., ROBISON, E.L. and MALIK, A.B. (1994). Touch-tone data entry for household surveys: research findings and possible applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 509-511.
- MAFFEO, C., FREY, W. and KALTON, G. (2000). Survey design and data collection issues in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.
- MAKLAN, D., and WAKSBERG, J. (1988). Within household coverage in RDD surveys. *Telephone Survey Methodology* (R. M. Groves, et al. - Eds.). New York: John Wiley and Sons, 51-69.
- MALAKHOFF, L.A., and APPEL, M.V. (1997). The development of a voice recognition prototype for field listing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 234-238.
- MASON, R.E., and IMMERMAN, F.W. (1988). Minimum cost sample allocation for Mitofsky-Waksberg random digit dialing. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 127-141.
- MASSEY, J.T. (1995). Estimating the response rate in a telephone survey with screening. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 673-677.
- MASSEY, J.T., and BOTMAN, S.L. (1988). Weighting adjustments for random digit dialed surveys. *Telephone Survey Methodology*, (R.M. Groves et al. - Eds.). New York: John Wiley and Sons, 143-160.
- MASSEY, J.T., O'CONNOR, D. and KROTKI, K. (1997). Response rates in random digit dialing (RDD) telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 707-712.
- MEEKS, R.L., LANIER, A.T., FECOSO, R.S. and COLLINS, M.A. (1998). Web-based data collection in national science foundation surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-353.
- MERKLE, D.M., BAUMAN, S.L. and LAVRAKAS, P.J. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1070-1075.
- MITCHELL, G.H., and ROGERS, E.M. (1958). Telephone interviewing. *Journal of Farm Economics*, 40, 743-747.
- MITOFSKY, W. (1970). Sampling of Telephone Households. Unpublished CBS memorandum.
- MOHADJER, L. (1988). Stratification of prefix areas for sampling rare populations. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 161-173.
- MULLET, G.M. (1982). The efficacy of plus-one dialing: self-reported status. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 575-576.
- NATHAN, G., and AFRAMIAN, N. (1996). An experiment with CATI in Israel. Paper presented at *InterCasic '96 Conference*, San Antonio, TX.
- NATHAN, G., and ELIAV, T. (1988). Comparison of measurement errors for telephone interviewing and home visits by misclassification models. *Journal of Official Statistics*, 4, 363-374.
- NICHOLLS, W.L., II (1983). CATI research and development at the Census Bureau. *Sociological Methods and Research*, 12, 191-198.
- NICHOLLS, W.L., II (1988). Computer-assisted telephone interviewing: A general introduction. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 377-386.
- NICOLAAS, G., LYNN, P. and LOUND, C. (2000). Random digit dialling in the UK: viability of the sampling method revisited. Paper presented at Fifth International Conference of Social Science Methodology, Koln.
- NORRIS, D.A., and PATON, D.G. (1991). Canada's General Social Survey: five years of experience. *Survey Methodology*, 17, 227-240.
- NTIA (2000). *Falling Through the Net, Toward Digital Inclusion*. Washington DC: National Telecommunications and Information Administration.
- NUSSER, S., and THOMPSON, D. (1998). Web-based survey tools. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 951-956.
- OAKES, R.H. (1954). Differences in responsiveness in telephone versus personal interviews. *Journal of Marketing*, 19, 169.
- OKSENBERG, L., and CANNEL, C. (1988). Effects of interviewer vocal characteristics on nonresponse. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 257-269.
- OLDENDICK, R.W., BISHOP, G.F., SORENSON, S.B. and TUCHFARBER, A.J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, 4, 307-318.
- OLDENDICK R.W., and LINK, M.W. (1994). The answering machine generation: who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58, 264-273.
- OFTEL (1999). *Homes Without a Fixed Line Phone - Who Are They?* [<http://www.oftel.gov.uk/publications/research/unph0400.htm>].
- OFTEL (2000). *Consumers' use of Internet*. [<http://www.oftel.gov.uk/publications/research/int1000.htm>]
- O'REILLY, J.M., HUBBARD, M.L., LESSLER, J.T., BIEMER, P.P. and TURNER, C.F. (1994). Audio and video computer assisted self-interviewing: preliminary tests of new technologies for data collection. *Journal of Official Statistics*, 10, 197-214.

- O'ROURKE, D., and BLAIR, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428-432.
- PALIT, C.D. (1980). A microcomputer based computer assisted interviewing system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 243-244.
- PALIT, C.D. (1983). Design strategies in RDD sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 627-629.
- PALIT, C.D., and BLAIR, J. (1986). Some alternatives for the treatment of first phase telephone numbers in a Waksberg-Mitofsky RDD sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 363-369.
- PALIT, C.D., and SHARP, H. (1983). Microcomputer-assisted telephone interviewing. *Sociological Methods and Research*, 12, 169-189.
- PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.
- PAYNE, S. L. (1956). Some advantages of telephone surveys. *Journal of Marketing*, 20, 278-281.
- PERNEGER, T.V., MYERS, T.L., KLAG, M.J. and WHELTON, P.K. (1993). Effectiveness of the Waksberg telephone sampling method for the selection of population controls. *American Journal of Epidemiology*, 138, 574-584.
- PERONE, C., MATRUNDOLA, G. and SOVERINI, M. (1999). A quality control approach to mobile phone surveys; the experience of Telecom Italia Mobile. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 180-187.
- PERRY, J. B. (1968). A note on the use of telephone directories as a sample source. *Public Opinion Quarterly*, 32, 691-695.
- PHIPPS, P.A., and TUPEK, A.R. (1991). Assessing measurement errors in a touchtone recognition survey. *Survey Methodology*, 17, 15-26.
- PIAZZA, T. (1993). Meeting the challenge of answering machines. *Public Opinion Quarterly*, 57, 219-231.
- POTTER, F.J., MCNEILL, J.J., WILLIAMS, S.R. and WAITMAN, M.A. (1991). List-assisted RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 117-122.
- POTTHOFF, R.F. (1987a). Some generalizations of the Mitofsky-Waksberg technique of random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- POTTHOFF, R.F. (1987b). Generalizations of the Mitofsky-Waksberg technique for random digit dialing: some added topics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 615-620.
- POYNTER, R. (2000). *We've Got Five Years*. London: Association for Survey Computing meeting on Survey Research on the Internet.
- RAMOS, M., SEDIVI, B.M. and SWEET, E.M. (1998). Computerized self-administered questionnaires. *Computer Assisted Survey Information Collection* (M.P. Couper, et al. - Eds.). New York: John Wiley and Sons, 389-408.
- RANTA-AHO, M., and LEPPINEN, A. (1997). Matching telecommunication services with user communication needs. *Proceedings of the International Symposium on Human Factors in Telecommunications*, (K. Nordby and L. Grafisk - Eds.). Oslo, Norway, 401-408. [<http://www.comlab.hut.fi/hft/publications/matcharticle.pdf>].
- RICH, C.L. (1977). Is random digit dialing really necessary? *Journal of Marketing Research*, 14, 300-305.
- ROGERS, T.F. (1976). Interviews by telephone and in person: quality of responses and field performance. *Public Opinion Quarterly*, 40, 51-65.
- ROGERS, S.M., MILLER, H.G., FORSYTH, B.H., SMITH, T.K. and TURNER, C.F. (1996). Audio-CASI: the impact of operational characteristics on data quality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1042-1047.
- ROMUALD, K.S., and HAGGARD, L.M. (1994). The effect of varying the respondent selection script on respondent self-selection in RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1299-1304.
- ROSEN, R.J., MANNING, C.D. and HARRELL, L.J., Jr. (1998). Web-based data collection in the current employment statistics survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.
- ROSLOW, S., and ROSLOW, L. (1972). Unlisted phone subscribers are different. *Journal of Advertising Research*, 7, 4, 35-38.
- ROUQUETTE, C. (2000). La percée du téléphone portable et d'Internet. *INSEE Première No. 200*. INSEE, Paris. [http://www.insee.fr/fr/ffc/docs_ffc/ip700.pdf].
- ST. CLAIR, J., and MUIR, J. (1997). Household adoption of digital technologies. *Year Book Australia 1997*. Canberra: Australian Bureau of Statistics.
- SALMON, C.T., and NICHOLS, J.S. (1983). The next birthday method of respondent selection. *Public Opinion Quarterly*, 47, 270-276.
- SCHEUREN, F., and PETSKA, T. (1993). Turning administrative systems into information systems. *Journal of Official Statistics*, 9, 109-119.
- SCHMIEDESKAMP, J. W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.
- SEBOLD, J. (1988). Survey period length, unanswered numbers, and nonresponse in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 247-256.
- SHANKS, J.M. (1983). The current status of computer assisted telephone interviewing: recent progress and future prospects. *Sociological Methods and Research*, 12, 119-142.
- SHANKS, J.M., NICHOLLS, W.L., II and FREEMAN, H.E. (1981). The California Disability Survey: design and execution of a computer-assisted telephone study. *Sociological Methods and Research*, 10, 123-140.
- SHAPIRO, G. M., BATTAGLIA, M. P., HOAGLIN, D. C., BUCKLEY, P., and MASSEY, J. T. (1996). Geographical variation in within-household coverage of households with telephones in an RDD survey. *Proceedings of the Section on Survey Research Methods*, 491-496.

- SHURE, G.E., and MEEKER, R.J. (1978). A mini-computer system for multi-person computer-assisted telephone interviewing. *Behavior Methods and Instrumentation*, (April) 196-202.
- SMITH, C., and FRAZIER, E. L. (1993). Comparison of traditional and modified Waksberg. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 926-931.
- SQUIRE, P. (1988). Why the 1936 Literary digest poll failed. *Public Opinion Quarterly*, 52, 125-133.
- STATISTICS CANADA (2000). *Selected Dwelling Characteristics and Household Equipment*. Income Statistics Division. [<http://www.statcan.ca/english/Pgdb/People/Families/famil09b.htm>].
- STOCK, J. S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 3, 50-51.
- STOKES, L., and YEH, M.-Y. (1988). Searching for causes of interviewer effects in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 357-376.
- SUDMAN, S. (1966). New uses of telephone methods in survey research. *Journal of Marketing Research*, 3, 107-120.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- SUDMAN, S. (1978). Optimum cluster designs within a primary unit using combined telephone screening and face-to-face interviewing. *Journal of the American Statistical Association*, 73, 300-304.
- SURVEY RESEARCH CENTER (2000). *Sample Design for Household Telephone Surveys: A Bibliography 1949-1996*. College Park, MD: University of Maryland. [<http://www.bsos.umd.edu/src/sampbib.html>].
- SURVEY SAMPLING INC. (1998). Random digit samples – part 1. [[http://www.ssisamples.com/ssi.x2o\\$ssi_gen.search_item?id=119](http://www.ssisamples.com/ssi.x2o$ssi_gen.search_item?id=119)].
- STATISTICS NETHERLANDS (1987). *Automation in Survey Processing*. Voorburg/Heerlen: Netherlands Central Bureau of Statistics (CBS Select 4).
- SYKES, W.M., and COLLINS, M. (1987). Comparing telephone and face-to-face interviewing in the United Kingdom. *Survey Methodology*, 13, 15-28.
- SYKES, W.M., and COLLINS, M. (1988). Effects of mode of interview: experiments in the UK. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 301-320.
- THORNBERRY, O.T. JR., and MASSEY, J.T. (1978). Correcting for undercoverage bias in random digit dialed National Health Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 224-229.
- THORNBERRY, O.T. JR., and MASSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 654-659.
- THORNBERRY, O.T. JR., and MASSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 25-49.
- TORTORA, R.D. (1985). CATI in an agricultural statistics agency. *Journal of Official Statistics*, 1, 301-314.
- TOURANGEAU, R., and SMITH, T.W. (1998). Collecting sensitive information with different modes of data collection. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et al. - Eds.). New York: John Wiley and Sons, 431-453.
- TRAUGOTT, M.W., GROVES, R.M. and LEPKOWSKI, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, 51, 522-539.
- TREWIN, D., and LEE, G. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (R.M. Groves, et al. - Eds.). New York: John Wiley and Sons, 3-24.
- TROLD AHL, V.C., and CARTER, R.E. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71-76.
- TUCKEL, P.S., and FEINBERG, B.M. (1991) The answering machine poses many questions for telephone survey researchers. *Public Opinion Quarterly*, 55, 200-217.
- TUCKEL, P.S., and O'NEILL, H. (1996). New technology and nonresponse bias in RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 889-894.
- TUCKEL, P., and SHUKERS, T. (1997). The effect of different introductions and answering machine messages on response rates. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1047-1051.
- TUCKER, C., CASADY, R. and LEPKOWSKI, J. (1992). Sample allocation for stratified telephone sample designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 291-296.
- TURNER, C.F., FORSYTH, B.H., O'REILLY, J.M., COOLEY, P. C., SMITH, T.K., ROGERS, S.M., and MILLER, H.G. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et al. - Eds.). New York: John Wiley and Sons, 455-473.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- WAKSBERG, J. (1983). A note on 'Locating a special population using random digit dialing'. *Public Opinion Quarterly*, 47, 576-578.
- WAKSBERG J. (1984). *Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling*. Unpublished Westat memorandum.
- WAKSBERG, J., BRICK, J.M., SHAPIRO, G., FLORES-CERVANTES, I. and BELL, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-718.
- WERKING, G., TUPEK, A. R., and CLAYTON, R. L. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 4, 349-362.
- WHITE, A. A. (1983). Response rate calculation in RDD telephone health surveys: current practices. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 277-282.

- WHITMORE, R. W., MASON, R. E., and HARTWELL, T. D. (1985). Use of geographically classified telephone directory lists in multi-mode surveys. *Journal of the American Statistical Association*, 80, 842-844.
- WILSON, P., BLACKSHAW, N., and NORRIS P. (1988). An evaluation of telephone interviewing on the British Labour Force Survey. *Journal of Official Statistics*, 4, 385-400.
- WINTER, D. L. S., and CLAYTON, R. L. (1990). Speech data entry: results of the first test of voice recognition for data collection. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 387-392.
- WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly*, 36, 105-108.
- XU, M. BATES, B.J., and SCHWEITZER, J.C. (1993). The impact of messages on survey participation in answering machine households. *Public Opinion Quarterly*, 57, 232-237.

Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design

AVINASH C. SINGH, BRIAN KENNEDY and SHIYING WU¹

ABSTRACT

We consider the regression composite estimation introduced by Singh (1994, 1996; termed earlier as "modified regression composite" estimation), a version of which (suggested by Fuller 1999) has been implemented for the Canadian Labour Force Survey (CLFS) beginning in January 2000. The regression composite (rc) estimator enhances the generalized regression (gr) estimator used earlier for the CLFS and the well known Gurney-Daly *ak*-composite estimator in several ways. The main features of the rc-estimator are: (a) it considerably improves the efficiency of level and change estimates for key study variables resulting into less volatile estimate series; (b) it is calculated like the gr-estimator as a calibration estimator such that all the usual poststratification controls used in gr as well as the new controls corresponding to correlated variables from the previous time point are met; and (c) it respects the internal consistency of estimators without having to calculate part estimates differently as residuals. The main innovations used in rc-class of estimators entail: (a) using the idea of working covariance matrix in estimating functions as an alternative to superpopulation modeling for defining regression coefficients for the predictors in the gr-estimator, (b) treating random controls (the ones based on the key correlated variables from past) as fixed, while computing the regression coefficients, similar to two-phase estimation, and motivated from the working covariance idea, and (c) that of the use of micro-matching to obtain previous time point's micro-level auxiliary information for realizing higher correlation with the present time point's study variables. As a by product, a new version of the *ak*-estimator which uses the micro-matching based predictors from past rather than the traditional macro-level is recommended in the interest of higher efficiency gains. The paper also presents an interesting heuristic justification of the smoothness feature of composite estimates using the amortization idea. Empirical results based on the Ontario 1996 CLFS data are presented for comparison of various estimators.

KEY WORDS: Generalized regression; Modified regression; Estimating functions; Regression calibration.

1. INTRODUCTION

In the case of repeated surveys with partially overlapping samples, it is well known (see, *e.g.*, Cochran 1977, Ch. 12) that estimates of level at a point in time and change between two time points can be improved by regressing the usual cross-sectional estimator (typically regression or simply Horvitz-Thompson) on the new predictors provided by the correlated observations on the overlapping subsample from the previous time point. Such methods of estimation belong to the class of composite estimation, and a simple version of which known as the *k*-composite estimator was proposed some time ago by Hansen, Hurwitz and Madow (1953), and examined further by Rao and Graham (1964), Binder and Hidiroglou (1988) provide an excellent review of the literature on estimation with repeated surveys. Note that there is an associated loss of efficiency in estimates aggregated over several time points due to increased positive correlation between composite estimates of successive time points. This is, however, probably a small price to pay because it is not the aggregate, but the level and change estimates that need more precision. The *ak*-composite estimator of Gurney and Daly (1965) provides an improved version of the *k*-composite estimator by reducing the variance further, an alternative simpler justification of which was provided by Wolter (1979).

The composite estimator considered in this paper was developed in the context of the Canadian Labour Force Survey (CLFS). The CLFS is a monthly survey that follows a rotating panel design with six panels. In any two consecutive months, five sixth of the households form the overlapping sample. It was in January 2000 that the CLFS started using a version (suggested by Fuller 1999) of the composite estimators introduced by Singh (1994, 1996) termed originally as "modified regression composite" estimators, which will be referred to in this paper as simply "regression composite" or rc-estimators. Before January 2000, CLFS used the generalized regression (gr) estimators of Cassel, Särndal, and Wretman (1976) and Särndal (1980) which were based on only cross-sectional (*i.e.*, present month's) data. It has long been felt that the estimator for CLFS could be improved using the composite estimation idea in the sense that estimates of level and change would be more efficient, and hence the resulting series would be more stable, *i.e.*, less volatile. There are four goals that the rc-estimator attempts to meet in modifying the gr-estimator:

- (i) It should considerably increase the efficiency of level and change estimates so that the estimate series becomes smoother or less volatile.
- (ii) It can be computed as a calibration estimator like the gr-estimator so that the existing estimation software system can be used with little modification,

¹ Avinash C. Singh and Shiyong Wu, Statistical Research Division, Research Triangle Institute, Research Triangle Park, N.C. 27709-2194, U.S.A.; Brian Kennedy, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

- (iii) The final calibrated weights should continue to satisfy the usual demographic and geographic controls used in the gr-estimator in addition to some new controls based on past month's variables, and
- (iv) The estimator should have the internal consistency property in that the part composite estimators add up to the whole, e.g., estimates for Employed (E), Unemployed (U), and Not in the Labour Force (N) should add up to the total eligible population in the domain of interest.

The *ak*-estimator was studied by Kumar and Lee (1983) in the context of CLFS, and it was found that it didn't give substantial gains in efficiency as required by goal (i). The goal (iv) was, of course, known to be not satisfied by the *ak*-estimator because the (optimal) coefficients *a* and *k* used for combining several present month's estimators (in fact three of them, one is the usual estimator based on the present month, and the other two are built on predictors from the past month) turn out to be specific to the characteristic such as E. A solution (although rather undesirable) is to designate one of the components as least important (say, N) and then obtain its estimate as a residual. The goals (ii) and (iii) can, however, be met by the *ak*-composite weighting suggested by Fuller (1990), and studied for the US Current Population Survey context by Lent, Miller, and Cantwell (1994, 1996). The goal (ii) is important especially for unplanned study variables for which the coefficients (*a, k*) are not known in advance. The rc-estimator meets all the four goals, in particular the goals (i) and (iv), by making use of the following three innovations:

- (i) The design-based estimation in the presence of correlated predictors can be cast in an estimating functions framework as defined by Godambe and Thompson (1989), and then use the idea of working covariance matrix as in Liang and Zeger (1986) to obtain an alternative to the superpopulation modelling to compute regression coefficients. The resulting regression estimates, like gr, are only suboptimal under the design randomization.
- (ii) The previous month's full sample composite estimates used as regression controls for present month's estimation can be treated as fixed using the working covariance idea for computational simplicity without violating the design consistency property. For variance estimation, the extra variation due to random controls should, of course, be accounted for.
- (iii) Using micro-matching of the present month's overlapping subsample with the previous month, information about key study variables from the previous month is augmented to the present month's data. These now serve as additional covariates deemed to be highly correlated with the present month's study variable.

These innovations allow for computation of all estimates using the gr-system, thus avoiding the need of having to compute parts of estimates as residuals in the interest of internal consistency. The feature of micro-matching gives rise to desired gains in efficiency. In practice, it would often be the case that some of the present month's respondents in the overlapping sample were nonrespondents in the previous month, and so imputation might be necessary. In the case of CLFS, this is a small fraction, and the Hot Deck method with donor classes defined by demographic, geographic (subprovincial economic regions), type of area (rural/urban), present month's employment status, and industry group is used to fill in the missing values. It may be noted that sometimes imputation may be necessary not due to nonresponse at the previous time point, but due to the household's move. Assuming that on the average, households that move in the dwellings sampled at the present time *t* are similar to the households that move out at *t*, then even though movers may have different employment characteristics than nonmovers, the imputation for movers is not expected to introduce any new bias as current month's employment status among other covariates is taken into account.

In the concluding section 6, a method is suggested to diagnose the impact of this imputation. This impact may be serious for surveys with high fraction of previous month's missing values for the present month's respondents in the overlapping subsample. A possibly simple way out would be to redesign the questionnaire so that the interviewer is prompted by the instrument CATI software (computer assisted telephone interviewing commonly used now-a-days) while administering the interview in second or later months, whether the respondent was nonrespondent at the previous month. If so, then the interviewer administers a rather short supplementary questionnaire in order to elicit the respondent's employment status for the previous month. This idea is similar to the method suggested by Hansen-Hurwitz-Madow for completely nonoverlapping repeated surveys, but each respondent is asked questions for the present as well as the previous time point, see Cochran (1977, page 355).

The organization of this paper is as follows. Section 2 presents a heuristic motivation using the amortization idea of why composite estimation, in general, is expected to provide desired smoothing of the estimate series. Section 3 defines various estimators, and discusses their computation via the gr-system. A new version of the *ak*-estimator, denoted by *ak**, is also proposed. The estimator uses predictors from previous month based on micro-matching, and is expected to give high gains in efficiency. Section 4 considers variance estimation by the currently used method of jackknife. An empirical comparison of the estimators is presented in section 5 using the Ontario 1996 CLFS data. Finally section 6 contains concluding remarks.

2. SERIES SMOOTHING BY COMPOSITE ESTIMATION: HEURISTICS

In this section, we present an interesting heuristic justification (based on the amortization idea rather than the shrinkage) of why smoothing of the estimate series is expected by composite estimation. (Using only the shrinkage idea, the series can be smoothed but it may not cross the original series often enough. With amortization, however, the left-over part after shrinkage is accounted for gradually over time, thus allowing for the smoothed series to cross the original one more often.) Consider the panel rotation scheme similar to that of the CLFS and let γ denote the fraction of the panels rotated out; in the case of CLFS, γ is 1/6. Denote the cross-sectional estimator (typically gr) at time t based on all panels, *i.e.*, the full sample, by F_t , the estimator based on only the birth (*i.e.*, rotate-in) panel by B_t , and the one based on nonbirth panels (*i.e.*, the subsample at t overlapping with the past sample at $t-1$) be \bar{B}_t . Similarly, denote the estimator based only on the death (*i.e.*, rotate-out) panel by D_t , and the one based on nondeath panels (*i.e.*, the subsample at $t-1$ overlapping with the present sample at t) be \bar{D}_t . We have

$$F_t = \gamma B_t + (1 - \gamma) \bar{B}_t \quad (2.1a)$$

$$F_{t-1} = \gamma D_{t-1} + (1 - \gamma) \bar{D}_{t-1}. \quad (2.1b)$$

Suppose, the series $\{F_t\}$ is too volatile, and we wish to smooth it. In the following it is assumed that there is no rotation group bias (Bailar 1975), *i.e.*, different rotation groups have the same expected value. Thus F_t is unbiased but may be unstable. This set-up is the traditional one for composite estimation in which different unbiased estimates are combined optimally to get a more efficient estimate. However, see the discussion at the end of this section for an alternative perspective on composite estimation in the presence of rotation group bias. Now denote the smoothed series by $\{C_t\}$, and consider the identity:

$$F_t = C_{t-1} + (F_t - F_{t-1}) + (F_{t-1} - C_{t-1}). \quad (2.2)$$

The above relation can be interpreted as follows. The estimate C_{t-1} at $t-1$ is adjusted by the fluctuation $(F_t - F_{t-1})$ at the next time point t in the F -series, and the existing gap $(F_{t-1} - C_{t-1})$ at the time point $t-1$. If we define C_t after full adjustments for these two differences, then C_t would be the same as F_t and there would be no smoothing of the F -series. This suggests that the adjustments for the differences $(F_t - F_{t-1})$ and $(F_{t-1} - C_{t-1})$ should be accounted for only partially as C -series moves from C_{t-1} to C_t . The remaining portions of the differences should be amortized gradually over future time points. All these adjustments should be done without affecting unbiasedness of the estimator C_t . The difference $(F_{t-1} - C_{t-1})$ is zero in expectation assuming unbiasedness of C_{t-1} and F_{t-1} (which is so under the assumption of no rotation group bias) and therefore amortizing parts of it

would not affect unbiasedness of future estimates C_t . However, the difference $F_t - F_{t-1}$ is not zero in expectation, and care should be exercised in amortizing part of this difference. Observe that

$$F_t - F_{t-1} = (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) + \gamma(\bar{D}_{t-1} - D_{t-1}). \quad (2.3)$$

The first term on the RHS is the change estimate based on common panels, while the second and third terms represent birth and death effects at t and $t-1$ respectively. The last two terms are zero functions (*i.e.*, are zero in expectation) but the first one is not. (Fortunately, the first term is expected to be stable as it is a difference of two highly correlated estimates.) Therefore, it is the second and third terms that should be amortized. Now, write (2.2) as

$$\begin{aligned} F_t &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) \\ &\quad + [\gamma(\bar{D}_{t-1} - D_{t-1}) + (F_{t-1} - C_{t-1})] \\ &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) \\ &\quad + [(\bar{D}_{t-1} - F_{t-1}) + (F_{t-1} - C_{t-1})] \\ &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \gamma(B_t - \bar{B}_t) + (\bar{D}_{t-1} - C_{t-1}). \end{aligned} \quad (2.4)$$

and define two amortization factors δ_{1t} , δ_{2t} between 0 and 1, and then define the smoothed series $\{C_t\}$ as

$$C_t = C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + \delta_{1t} \gamma(B_t - \bar{B}_t) + \delta_{2t} (\bar{D}_{t-1} - C_{t-1}). \quad (2.5)$$

The term with δ_{1t} in (2.5) represents shrinkage of the birth effect at t which C_t tries to account for, while the term with δ_{2t} refers approximately to shrinkage of the death effect at the past time $(t-1)$ which C_t tries to make up for the present time t . Also, it would be desirable to set $\delta_{2t} < \delta_{1t}$ in order for the series $\{C_t\}$ to track $\{F_t\}$ better so that they have similar trend over time, *i.e.*, give more importance to the current birth effect than the past death effect. (In fact, a rigorous justification under fairly general conditions of why one should set $\delta_{2t} < \delta_{1t}$ comes from optimality considerations in which variance of C_t is minimized to obtain the best linear combination of three unbiased estimators, F_t , $C_{t-1} + \bar{B}_t - \bar{D}_{t-1}$, and $F_t + C_{t-1} - \bar{D}_{t-1}$ of the present month's population total; see (2.8) at the end of this section for the actual expression.) Now, to see the connection with the well known composite estimates defined in the next section, define $0 < a_t, b_t < 1$, so that $\delta_{1t} = 1 - b_t$, $\delta_{2t} = 1 - b_t - a_t$. We have

$$\begin{aligned} C_t &= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t) \gamma(B_t - \bar{B}_t) \\ &\quad + (1 - b_t - a_t) (\bar{D}_{t-1} - C_{t-1}). \end{aligned} \quad (2.6)$$

It is interesting to note that if $b_t = 0$, there would be no dampening of the birth effect, and the C -series is expected to be closer to F -series, *i.e.*, there is less smoothing and the two would cross each other more often. If $a_t = 0$, the past

effect represented by $(\bar{D}_{t-1} - C_{t-1})$ is dampened less. This would imply more smoothing of the F -series, and the two series are expected to cross each other less frequently. Finally, if $a_t, b_t > 0$, then the behaviour of the C -series relative to the F -series would be somewhere in the middle. Moreover, if b_t is high (close to 1), there would be quite a bit of smoothing of the F -series because there is high amortization of both the birth and death effects. In these situations, one would expect sustained gaps between F and C series over time before they cross each other. Notice that parts of the term $\gamma(B_t - \bar{B}_t)$ that get amortized over $t, t+1, \dots$ decrease as t increases. They are given by $b_t \gamma(B_t - \bar{B}_t)$, $(b_{t+1} + a_{t+1}) b_t \gamma(B_t - \bar{B}_t)$, \dots . Similarly, the amortized parts of $(\bar{D}_{t-1} - C_{t-1})$ are

$$(b_t + a_t)(\bar{D}_{t-1} - C_{t-1}), (b_{t+1} + a_{t+1})(b_t + a_t)(\bar{D}_{t-1} - C_{t-1}), \dots$$

Clearly, when b_t is large, it will take several time points for completing the amortization. However, as explained earlier, this would not introduce bias because the effects being amortized are zero functions under the assumption of no rotation group bias.

The expression (2.6) can be cast into a more familiar expression of the composite estimator as follows:

$$C_t = C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t)(F_t - \bar{B}_t) + (1 - b_t)(\bar{D}_{t-1} - C_{t-1}) + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7a)$$

$$= C_{t-1} + (\bar{B}_t - \bar{D}_{t-1}) + (1 - b_t)(F_t - \bar{B}_t + \bar{D}_{t-1} - C_{t-1}) + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7b)$$

$$= F_t + b_t[C_{t-1} - (F_t + \bar{D}_{t-1} - \bar{B}_t)] + a_t(C_{t-1} - \bar{D}_{t-1}) \quad (2.7c)$$

$$= F_t + (b_t + a_t)(C_{t-1} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a_t(F_t - \bar{B}_t). \quad (2.7d)$$

The expression (2.7d) coincides with the ak -estimator (see next section) when $a_t = a$ and $b_t + a_t = k$. In practice, the values of a_t and b_t can be determined optimally or suboptimally using regression (see next section). The partial regression coefficients a_t, b_t satisfy $0 < a_t < b_t < 1$ in general, because the direct estimator F_t is expected to be more positively correlated with the predictor $F_t + (\bar{D}_{t-1} - \bar{B}_t)$, i.e., $\bar{D}_{t-1} + \gamma(B_t - \bar{B}_t)$ than with the predictor \bar{D}_{t-1} ; both predictors being unbiased estimates, like C_{t-1} , of the population total parameter at the previous time point $t-1$. It follows from (2.7c) that the estimator C_t can be written as a linear combination of the three unbiased estimators mentioned earlier, and is given by

$$C_t = (1 - b_t - a_t)F_t + b_t(C_{t-1} + \bar{B}_t - \bar{D}_{t-1}) + a_t(F_t + C_{t-1} - \bar{D}_{t-1}). \quad (2.8)$$

The above heuristic motivation corresponds to the variance reduction considerations under the assumption of no rotation group bias when combining three unbiased estimators of the population total at t . In the presence of rotation group bias, however, all the three estimators become biased with possibly different magnitude and direction, and what composite estimation does is to adjust each one of them so that the adjusted value for each is equal to a common value given by the composite estimator. (For example, in the case of two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ , the linear combination $\lambda \hat{\theta}_1 + (1 - \lambda) \hat{\theta}_2$ can be written as $\hat{\theta}_2 + \lambda(\hat{\theta}_1 - \hat{\theta}_2)$ or $\hat{\theta}_1 + (1 - \lambda)(\hat{\theta}_2 - \hat{\theta}_1)$ implying that the two original estimators are adjusted appropriately to converge to a common value.) The relative weight in combining the three estimators depends on the criterion of minimum variance. Ideally, it should be based on the minimum MSE criterion, but it is hard to get a handle on bias because it can't be estimated. Clearly the composite estimator is not bias free, and it can only be speculated that the overall bias of the estimator is reduced by compositing. Similarly if, instead, a suboptimal regression is used in constructing the composite estimator (as in rc-estimation, see the next section), then what composite estimation does is to adjust the sampling weights in the full sample (which are generally gr-weights) so that $F_t - (\bar{B}_t - \bar{D}_{t-1})$, and \bar{D}_{t-1} with adjusted weights become equal to C_{t-1} ; the C_{t-1} serve as new controls in the calibration step. This is another way of adjusting the three estimators to a common value, but again bias of the resulting composite estimator remains unknown. The above discussion of two perspectives on composite estimation has some similarity with the dual property of poststratification in terms of both variance and (coverage) bias reduction, see Singh and Folsom (2000).

3. COMPOSITE ESTIMATORS: NEW AND OLD

We start with the cross-sectional estimator at time t of the total $\tau_y(t)$ defined as gr, which is given by

$$\hat{\tau}_{y(\text{gr})}(t) = \sum_{k \in s(t)} y_k(t) w_{\text{gr}}(t, k), \quad (3.1)$$

$$w_{\text{gr}}(t, k)$$

$$= d(t, k) \left[1 + x_k(t)' (X(t)' \Delta(t) X(t))^{-1} (\tau_x(t) - \hat{\tau}_x(t)) \right], \quad (3.2)$$

where $d(t, k)$'s are the initial design weights adjusted for nonresponse, $x_k(t)$ is a p -vector of covariates used for calibration (or poststratification), $X(t)$ is the $n(t) \times p$ matrix of x -observations, $n(t)$ is the sample size, $\Delta(t)$ is $\text{diag}(d(t, k))$, $\tau_x(t)$ is the known p -vector of calibration controls, and $\hat{\tau}_x(t)$ is the corresponding vector of expansion estimates based on d -weights. In terms of the notation F_t, \bar{B}_t , and B_t of the previous section, F_t here can be taken as the gr-estimator (3.1), and \bar{B}_t is gr-estimator based on nonbirth panels given by

$$\bar{B}_t = (1 - \gamma)^{-1} \sum_{k \in s(t|t-1)} y_k(t) w_{gr}(t, k), \quad (3.3)$$

where $s(t|t-1)$ is the subsample at t matched with the sample at $t-1$. The estimator B_t is also a gr-estimator, and is given by

$$B_t = \gamma^{-1} \sum_{k \in s(t) - s(t|t-1)} y_k(t) w_{gr}(t, k), \quad (3.4)$$

where the sum is over the subsample defined by the birth panel at t .

The ak -composite estimator uses the macro-level past information for the new predictors, and can be defined as

$$\begin{aligned} C_{t(ak)} &= F_t + k(C_{t-1(ak)} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a(F_t - \bar{B}_t) \\ &= F_t + (k-a)(C_{t-1(ak)} - \bar{D}_{t-1} + \bar{B}_t - F_t) + a(C_{t-1(ak)} - \bar{D}_{t-1}). \end{aligned} \quad (3.5)$$

Here the coefficients a, k for level estimation are obtained by optimally regressing F_t on the two predictor zero functions, based on the past information, namely, $C_{t-1(ak)} - (F_t + \bar{D}_{t-1} - \bar{B}_t)$, and $C_{t-1(ak)} - \bar{D}_{t-1}$. Thus, a, k depend on the sample design as well as on the study variable y , in particular, they are not even the same for level and change estimates for the same y . For change estimation, $F_t - C_{t-1(ak)}$, and not F_t is regressed optimally on the above predictors. In practice, a, k are estimated by performing a grid search on the interval $(0,1)$ such that the variance of C_t is minimized. As mentioned earlier, typically a is smaller than k . In defining the above two new predictor zero functions based on past information, two estimators of $\tau_t(t-1)$ are first formed: one is \bar{D}_{t-1} based on the nondeath panels at $t-1$ (i.e., subsample at $t-1$ matched with the sample at t), and the other is $F_t + (\bar{D}_{t-1} - \bar{B}_t)$ which is the gr-estimator at time t adjusted for change from $t-1$ to t estimated from the common sample. Clearly, if there is no overlap in the panel design, then all the predictor zero functions become no longer meaningful resulting in no change in F_t by composite estimation. Similarly, if there is a complete overlap, then $\bar{B}_t = F_t$, and again there is no effect on F_t of composite estimation. This may at first seem counter-intuitive, because the past data (y_{t-1}) is correlated with the present (y_t) due to sample overlap. However, complete overlap amounts, in principle, to collecting a single sample of multivariate data on y with elements corresponding to y at different time points. Using this analogy, there is no room for improvement (in the design-based framework) as there is no larger sample with additional information. In the case of no overlap, additional information is there but it doesn't help as it is uncorrelated. Note, however, that at the first stage, psu's (primary sampling units) in CLFS remain common over several years before they are rotated out. Therefore, efficiency gains due to partial overlap are realized mainly from the reduction of the second stage variance component.

Furthermore, note that the estimator $C_{t(ak)}$ uses past information in the univariate sense in that for the study variable y , past information about only y_{t-1} is used. If new predictors based on several variables such as y_{t-1}, z_{t-1}, \dots from the past are also used for the study variable y , then the composite estimation becomes multivariate. However, the optimal choice of the (a, k) coefficients for the multivariate case can be quite cumbersome.

The rc-class of estimators is given by

$$\begin{aligned} C_{t(rc)} &= F_t + b_{t(rc)} \left(\tilde{C}_{t-1(rc)} - \bar{D}_{t-1}^* + \bar{B}_t - F_t \right) \\ &\quad + a_{t(rc)} \left(\tilde{C}_{t-1(rc)} - \bar{D}_{t-1}^* \right) \end{aligned} \quad (3.6)$$

where $\tilde{C}_{t-1(rc)}$ denotes the $t-1$ estimator for the study variable (y) after the $(t-1)$ -calibration weights are further calibrated to meet the controls used for poststratification by gr at time t . Thus $\tilde{C}_{t-1(rc)}$ is an estimate of the population total at t for the y -variable at $t-1$. The starred \bar{D}_{t-1}^* signifies that it is based on the subsample at t matched with the sample at $t-1$, but uses the gr-weights at t as the y values from $t-1$ are augmented to the sample at t by micro-matching. (Note that the estimator \bar{D}_{t-1}^* involves, in general, imputed values, and may suffer from bias due to imputation. For a diagnosis and adjustment for this bias, see section 6.) The coefficients $b_{t(rc)}$ and $a_{t(rc)}$ are computed similar to gr of (3.1); see below for more details. These coefficients are suboptimal unlike (a, k) . However, like (a, k) , they are y -specific, and in the case of multivariate they depend on the key set of study variables chosen from past for new controls, but they can be computed easily as they are suboptimal in nature. Thus with rc-estimation, it is fairly easy to introduce more predictors. The predictors $(C_{t-1} - \bar{D}_{t-1})$ and $(C_{t-1} - \bar{D}_{t-1} + \bar{B}_t - F_t)$ can be termed respectively as level-driven and change-driven as in Singh, Kennedy, Wu and Brisebois (1997). The reason for this is that not only the former is a difference of two level estimates, and the latter a difference of two change estimates, $(C_{t-1} - F_t)$ and $(\bar{D}_{t-1} - \bar{B}_t)$, but that the former tends to provide high efficiency gains in level estimation over what can be obtained in the presence of the latter, and similarly, the latter provides high efficiency gains in change estimation over what can be achieved in the presence of the former.

The idea of using the micro-level past information for the new predictors in rc-estimation can be applied to the ak -estimator, and thus a new estimator ak^* can be proposed.

$$\begin{aligned} C_{t(ak^*)} &= F_t + (k^* - a^*) \left(\tilde{C}_{t-1(ak^*)} - \bar{D}_{t-1}^* + \bar{B}_t - F_t \right) \\ &\quad + a^* \left(\tilde{C}_{t-1(ak^*)} - \bar{D}_{t-1}^* \right). \end{aligned} \quad (3.7)$$

The control $\tilde{C}_{t-1(ak^*)}$ denotes the $(t-1)$ calibration estimator for y after the ak^* -composite weights are further

calibrated to meet the controls used for poststratification by gr at t . (Here the ak^* -composite weights are similar to the ak -composite weights of Fuller (1990) where the composite estimators for a set of key y -variables serve as additional controls in the usual gr to obtain a set of final calibration weights. This allows for the ak -composite estimator to be computed as a calibration estimator.) The main differences between the various estimators defined above lie in the definition of regression coefficients (optimal vs. suboptimal), and that of the predictors (macro-level vs. micro-level use of past information). Special cases of the above composite estimators can be obtained as described in Singh, *et al.* (1997) by using only one of the two predictors. For $C_{t(ak)}$, if $a=0$ (i.e., only change-driven predictor is used), we get the well known k -composite estimator which can be termed as the $ak2$ -estimator in the present context. If $a=k$, i.e., only level-driven predictor is used, we get a new composite estimator $C_{t(ak1)}$ which can be termed as the $ak1$ -estimator. Similarly for $C_{t(ak*)}$, we get two more new composite estimators ak^*1 and ak^*2 . For $C_{t(rc)}$, with only level-driven predictor, we get the $rc1$ -estimator, termed earlier as MR1 in Singh and Merkouris (1995). With only change-driven predictors, we get the $rc2$ -estimator termed earlier as MR2 in Singh, *et al.* (1997).

As mentioned earlier, the rc -estimator is computed as a gr-estimator of (3.1), and therefore, it can be expressed as $\hat{y}_{y(rc)}(t) = \sum_{k \in s(t)} y_k(t) w_{rc}(t, k)$. The $X(t)$ -matrix is expanded to $n(t) \times (p + 2q)$ matrix $X(t)^*$ where $2q$ represents the number of new predictors, the factor 2 signifying the pair of level-driven and change-driven predictors. The (random) control totals $\tilde{C}_{t-1(rc)}$ corresponding to the key set of y -variables from $t-1$ selected for composite estimation are treated as fixed (during the computation of regression coefficients) like the other (nonrandom) gr-controls. Now, since the level-driven predictor can be written as

$$\begin{aligned} \tilde{D}_{t-1}^* &= (1-\gamma)^{-1} \sum_{k \in s(t-1)} y_k(t-1) w_{gr}(t, k) \\ &= \sum_{k \in s(t)} (1-\gamma)^{-1} y_k(t-1) 1_{k \in s(t-1)} w_{gr}(t, k) \end{aligned} \quad (3.8)$$

the column of the $X(t)^*$ -matrix corresponding to this predictor consists of $n(t)$ -values of $(1-\gamma)^{-1} y_k(t-1) 1_{k \in s(t-1)}$. Similarly the change-driven predictor can be written as

$$\begin{aligned} F_t + \tilde{D}_{t-1}^* - \tilde{B}_t &= \\ \sum_{k \in s(t)} \left(y_k(t) + (1-\gamma)^{-1} (y_k(t-1) - y_k(t)) 1_{k \in s(t-1)} \right) w_{gr}(t, k) \end{aligned} \quad (3.9)$$

and the corresponding column of the $X(t)^*$ matrix consists of the $n(t)$ -values of $y_k(t) + (1-\gamma)^{-1} (y_k(t-1) - y_k(t)) 1_{k \in s(t-1)}$. Once the $X(t)^*$ matrix is defined, the gr-system can be used to compute the calibration weights $w_{rc}(t, k)$ as in (3.2). Note that the calibration weights $w_{rc}(t, k)$ can be used for estimation of all study variables although they depend explicitly only on the key set of study variables chosen for the new predictors from correlated past information. Also

note that although the rc -estimator of (3.6) was defined as the gr -estimator plus regression-adjustments for the new predictors, computationally it is convenient to perform a gr -calibration on the design weights when all the old and new calibration controls are considered simultaneously. This way computation for the multivariate rc -estimator is not much different from the univariate rc -estimator. Alternatively, one could compute the rc -estimator as an adjusted gr as in (3.6), but the coefficients for the new predictors would be partial regression coefficients, and therefore do not have the standard form of the gr -coefficients.

Finally we note that with composite estimation, one would expect higher efficiency gains for change estimates ($C_t - C_{t-1}$ vs. $F_t - F_{t-1}$) than those for level estimates (C_t vs. F_t). To see this, consider a simple identity: $V(F_t - F_{t-1}) = V(F_t) + V(F_{t-1}) - 2\text{Cov}(F_t, F_{t-1})$. Typically $V(F_t) \approx V(F_{t-1}) = \sigma_{gr}^2$ (say), then the above can be reduced to $V(F_t - F_{t-1}) \approx 2\sigma_{gr}^2(1 - \rho_{gr})$. Similarly, $V(C_t - C_{t-1}) \approx 2\sigma_{rc}^2(1 - \rho_{rc})$. Thus the change efficiency is approximately the level efficiency times $(1 - \rho_{gr})/(1 - \rho_{rc})$. It follows that if the new predictors for composite estimation increase considerably the (positive) correlation between C_t and C_{t-1} , then the change efficiency will highly dominate the level efficiency.

4. VARIANCE ESTIMATION

The CLFS currently uses delete-one psu jackknifing to find variance of the gr -estimate. The method of jackknifing is valid (for cross-sectional surveys) if the psu-level estimates have identical mean and variance, and the psu selection can be treated as with replacement. When psu selection is without replacement the variance estimate becomes conservative if the (common) covariance between the psu-level estimates is negative. This is generally the case. For repeated surveys, a third condition that psu's are common (or connected) over time is needed. When this is the case the survey can be viewed as cross-sectional by treating the vector of observations (psu-level estimates) over time as a single observation collected at the conceptual end point in time. In the rotating panel design of the CLFS, psu's are not rotated out for a number of years, but the within psu units are rotated every six months. Each psu in the CLFS corresponds to a single panel which is either birth or non-birth. Note that to meet the conditions of jackknifing, it is not necessary that the same set of units be used to obtain psu-level estimates. The condition that psu-level estimates have common mean and variance within a stratum is reasonable on the grounds that the panel estimates have common mean and variance. For composite estimation, although birth and non-birth panels are treated differently, panel-level composite estimates should have identical mean and variance unconditionally on the panel assignment. This is so because the panels are assigned at random; a panel could have been birth with probability

$\gamma = 1/6$ and non-birth with probability $1 - \gamma = 5/6$. The resulting unconditional variance estimate will not be smaller than the one obtained conditionally on the panel assignment. Thus the method of jackknifing is expected to provide a conservative variance estimate in the CLFS context. Note that the above considerations for measures of uncertainty do not involve rotation group bias that may be present.

5. EVALUATION RESULTS

The numerical results are based on 1996 Ontario CLFS data, see Singh, *et al.* (1997). The auxiliary variables for *gr* are population counts corresponding to 16 age-sex groups, 11 economic regions, 10 census metropolitan areas, and 6 panels. Each panel control specifies 1/6 of the 15+ population. The new controls (30 in all) for *rc* corresponding to only change-driven predictors are: employed, unemployed and not in the labour force by age (young and old) by sex groups for a total of 12, employment by industry categories for a total of 16, and 2 employment by full/part time categories. In fact, these 30 new controls reduce to only 28 because of linear dependence. The multivariate *rc*-estimator involves these 28 extra controls, while the univariate *rc* involves just one extra control. The average relative efficiency shown in various tables is computed as the average variance of *gr* over 12 months of 1996 divided by the average variance of the composite estimator over 12 months.

5.1 Macro-level vs. Micro-level Predictors

For level-estimates, the correlation is computed between the current month level estimate (*i.e.*, F_t) and the predictor (*e.g.*, the level-driven $C_{t-1} - \bar{D}_{t-1}$ at the macro-level), whereas for the change estimate, it is computed between

$F_t - C_{t-1}$ and the predictors. The correlation is negative as expected because the estimate involving common panels is positively correlated with F_t but expressed with a negative sign in the predictor. Recall that the composite estimator used is the *ak* with macro-level and *ak** with micro-level predictors.

It is seen from Table 1 for the four key variables (employed, unemployed, employed in Trade, and employed in Transportation and Communication (TRCO)), for each of the level-driven and change-driven predictors, micro-level predictors outperform macro-level in terms of high correlation.

Between level- and change-driven predictors at the micro-level, change-driven is seen to out-perform level-driven. Similar results hold for other key variables. In view of these correlations, other evaluation results shown below pertain to only *ak2*, *ak*2*, and *rc2* versions of composite estimates. The *rc*-estimator with both level- and change-driven predictors was not included in the interest of keeping down the number of extra controls.

5.2 *ak* vs. *ak** vs. *rc* (Efficiencies Relative to *gr*)

Table 2 shows the optimal coefficients (*e.g.*, *k* for *ak2* estimator) and the corresponding relative efficiency over *gr*. The optimal coefficients were found via grid-search using the same 1996 data. (In practice, this should be based on past data). It is seen that the efficiency gains can be considerable as one moves from *ak* to *ak**. The optimal coefficients vary for level and change estimates. The last two columns under each of level and change estimates show the reduction in efficiency if level-optimal coefficients are used for change estimates and vice-versa. Level-optimal coefficients seem to perform quite well for change estimates, in contrast to a drop in efficiency of level estimates when change-optimal coefficients are used.

Table 1
Average Monthly Correlation between Composite Predictor and Estimates for Level and Change (Ontario, 1996)

Variable	Level				Change			
	Level-Driven Predictors		Change-Driven Predictors		Level-Driven Predictors		Change-Driven Predictors	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
Employed	-0.27	-0.35	-0.23	-0.45	-0.35	-0.49	-0.57	-0.84
Unemployed	-0.26	-0.35	-0.24	-0.33	-0.22	-0.40	-0.39	-0.53
Empl. Trade	-0.58	-0.55	-0.58	-0.65	-0.65	-0.73	-0.91	-0.96
Empl. TRCO	-0.58	-0.55	-0.60	-0.68	-0.63	-0.70	-0.92	-0.96

Table 2
Average Relative Efficiency of *ak* and *ak** over *gr* (Ontario, 1996)

Variable	Coeff				Eff (Level)		Eff (Change)		Eff (Level)	Eff (Change)
	Level Optimal		Change Optimal		Level Optimal		Change Optimal		Change Optimal	Level Optimal
	<i>ak</i>	<i>ak*</i>	<i>ak</i>	<i>ak*</i>	<i>ak</i>	<i>ak*</i>	<i>ak</i>	<i>ak*</i>	<i>ak*</i>	<i>ak*</i>
Employed	0.42	0.72	0.48	0.95	1.05	1.25	1.28	2.43	0.72	2.21
Unemployed	0.40	0.50	0.54	0.69	1.06	1.12	1.11	1.29	1.05	1.26
Empl. Trade	0.79	0.84	0.95	0.98	1.43	1.67	2.36	4.97	0.88	4.22
Empl. TRCO	0.84	0.87	0.95	0.98	1.59	1.88	3.60	7.59	1.11	6.51

Table 3 compares rc (univariate and multivariate) with ak^* . The possible values of $b_{(rc2)}$ coefficients over the 12 month-period for the univariate rc2 are summarized via mean, minimum and maximum. They can be compared with the corresponding optimal coefficients for ak^* . The rc-coefficients seem to provide a compromise and lie somewhere between level-optimal and change-optimal coefficient values. The rc-efficiencies for the change estimate are quite at par with those for ak^* but for level estimates, are somewhat lower. The efficiency gains at the aggregate level for which gr had controls are low but are high for domains without gr-controls.

Table 4 presents possible loss in efficiencies for estimates obtained as residuals in ak^* -estimation in the interest of internal consistency. It shows that caution should be exercised in practice when choosing variables for residual estimation or using compromise coefficient values in ak^* -estimation of components of an aggregate.

5.3 Change vs. Level Efficiencies of rc Over gr

Table 5 shows that the approximate relation (see section 3) between change and level efficiencies holds fairly well. It is seen that month-to-month correlation for rc-estimates for domains not having a corresponding population control in gr can be quite high compared to the correlation for gr.

This, in turn, yields a high factor by which change efficiency exceeds level efficiency.

5.4 Point Estimate and SE of Difference Between rc and gr

Table 6 shows monthly estimates (and SE of level and change estimates) for the variable (employed in trade at the Ontario level) for gr and rc. The corresponding values for the monthly difference (rc -gr) are also shown. It is seen that the differences between rc and gr are not significant in general. Efficiencies (not shown here) of annual average and quarterly estimates of rc and gr were also computed. As expected, due to serial correlation, there may be a loss in efficiency over gr. However in terms of the coefficient of variation, this is likely to be of no practical consequence.

5.5 Time Series of Level Estimates

Figures 1(a) and (b) show level estimates of employment for Ontario for the period 88-96 for gr and rc without and with seasonal adjustment. (The X11-ARIMA method was used.) Figures 2(a) and (b), show employment for the industry group "Trade". At the provincial level, aggregated over the industry group, there is similarity between gr and rc (seasonally adjusted or not) series because the gr-estimates have high precision to begin with. At the domain

Table 3
Average Relative Efficiency of rc over gr (Ontario, 1996)

Variable	Coeff			Eff (Change)					
	rc -univariate (level or change)			ak^*		rc		rc	
	Avg	Min	Max	Level	Change	(univariate)	(multivariate)	(univariate)	(multivariate)
Employed	0.88	0.81	0.90	0.72	0.95	1.05	1.05	2.39	2.46
Unemployed	0.60	0.53	0.65	0.50	0.69	1.12	1.12	1.31	1.33
Empl. Trade	0.96	0.94	0.98	0.84	0.98	1.17	1.22	4.98	5.07
Empl. TRCO	0.95	0.93	0.97	0.87	0.98	1.37	1.42	7.47	7.52

Table 4
Average Relative Efficiency of ak^* and rc over gr from Ontario, 1996 (Regular vs. Residual)

Variable		ak^* Coeff	Level		Change		
			Eff (ak^*)	Eff (rc)	ak^* Coeff	Eff (ak^*)	Eff (rc)
Agriculture	(regular)	0.91	2.55	2.32	0.97	4.88	5.22
Agriculture	(residual)	NA	0.63	2.32	NA	3.90	5.22
NILF	(regular)	0.74	1.26	1.07	0.95	1.96	2.01
NILF	(residual)	NA	1.21	1.07	NA	1.95	2.01

Table 5
Relation Between Level and Change Efficiencies for rc (multivariate) over gr (Ontario, 1996)

Variable	Change Eff	Level Eff	Change Eff/Level Eff	$(1-\rho_{gr})(1-\rho_{rc})$	ρ_{gr}	ρ_{rc}
Employed	2.46	1.05	2.34	2.65	0.77	0.91
Unemployed	1.33	1.12	1.19	1.21	0.50	0.59
Empl Trade	5.07	1.22	4.16	3.80	0.79	0.95
Empl TRCO	7.54	1.42	5.31	5.66	0.80	0.97

Table 6
 Monthly Point Estimates for gr and rc and Their Differences (Ontario, 1996)
 (Level and Change for Employment in Trade, Ontario, 1996)

Month	Type	gr		rc		rc-gr	
January	Level	886.5	(21.0)	858.9	(17.3)	-27.6	(23.0)
	Change	-25.8	(13.2)	-21.0	(5.6)	4.8	(11.4)
February	Level	906.5	(22.9)	867.9	(17.6)	38.6	(24.6)
	Change	20	(14.2)	9.0	(4.7)	-11.0	(12.5)
March	Level	927.1	(20.8)	874.1	(18.3)	-52.9	(23.1)
	Change	20.6	(13.3)	6.2	(4.7)	-14.4	(12.5)
April	Level	914.8	(20.3)	872.5	(17.7)	-42.3	(22.4)
	Change	-12.3	(13.4)	-1.6	(5.1)	10.7	(12.5)
May	Level	912.8	(18.9)	887.6	(17.0)	-25.1	(21.8)
	Change	-2.1	(13.0)	15.1	(5.7)	17.2	(11.6)
June	Level	908.1	(17.8)	888.6	(17.2)	-19.5	(21.5)
	Change	-4.7	(12.3)	0.9	(4.9)	5.6	(11.9)
July	Level	899.9	(18.1)	881.2	(17.7)	-18.7	(23.0)
	Change	-8.2	(12.8)	-7.4	(6.7)	0.8	(10.7)
August	Level	913.9	(16.9)	888.1	(18.3)	-25.8	(22.6)
	Change	14.0	(11.5)	6.9	(5.3)	-7.1	(10.3)
September	Level	886.6	(20.4)	876.4	(19.7)	-10.2	(23.1)
	Change	-27.3	(12.6)	-11.8	(6.3)	15.6	(11.1)
October	Level	898.6	(22.9)	889.3	(19.3)	9.3	(26.1)
	Change	12.1	(13.4)	12.9	(6.6)	0.9	(11.8)
November	Level	911.2	(20.3)	902.3	(19.3)	-8.9	(25.9)
	Change	12.6	(13.9)	13.0	(7.0)	0.4	(12.6)
December	Level	917.9	(20.5)	916.3	(19.0)	-1.5	(26.0)
	Change	6.7	(12.5)	14.0	(6.1)	7.4	(10.9)

Note: SEs are shown in parentheses.

level defined by Trade, however, the series are quite different. (Note that among numerous series that were examined, this particular series was chosen here to depict the extreme scenario for gaps between gr and rc series. For almost all other series, the two series crossed each other fairly often.) Since the gr-series is highly volatile, there is room for considerable smoothing by rc. Also note that because of expected high signal-to-noise ratio, seasonally adjusted rc series at the Trade-domain level looks considerably smoother than that for the gr-series; in fact, there is very little difference between with and without seasonally adjusted gr-series. It is also observed that there tends to be runs of consecutive periods where rc is either larger or smaller than gr. This is expected because of high values of the $b_{i(rc)}$ coefficients (Table 3), and high serial correlation in both series (see Table 5). Interestingly, turning points in the gr and rc series tend to occur at (approximately) same time points though they appear somewhat dampened with rc due to higher serial correlation in rc-series. It may be noted that the gap between the two series would have been smaller if controls for level-driven predictors were also included.

6. CONCLUDING REMARKS

The previously used gr-estimator in CLFS showed instability in change estimates and various domain level estimates. The rc-estimator provides smoother estimate series (which, in turn, renders change estimates more stable). The rc-method departs from the traditional *ak*-composite estimation in several ways, the main points being the use of micro-matching for collection of unit-level past information for common panels, and the use of regression calibration (like gr) to produce a set of final weights for use with all study variables. Three versions of rc were examined. Although this paper was mainly concerned with rc2, *i.e.*, with change-driven predictors (because of the desired resulting smoothness in estimate series), it was found (although not reported here) that level estimates of some key variables can be further improved (in comparison to rc2) by including corresponding level-driven predictors. Thus, in practice, a good strategy might be to use a mixture of mostly change-driven and some level-driven predictors.

The version of the rc-estimator currently implemented for CLFS was suggested by Fuller (1999), and can be expressed as

Figure 1(a) Employment in Ontario, actual

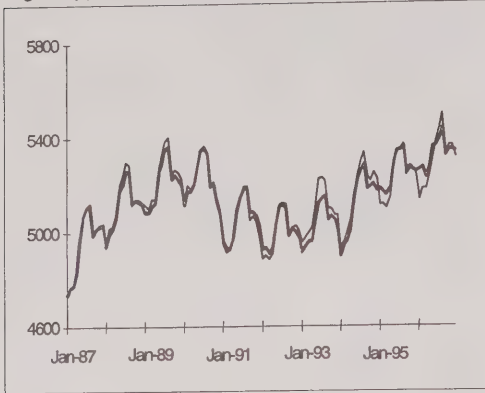


Figure 1(b) Employment, Ontario, seasonally adjusted

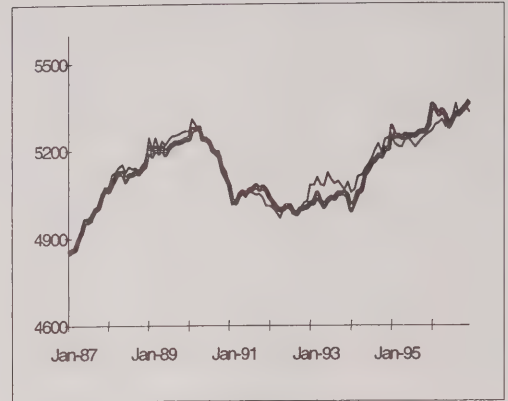


Figure 2(a) Employment in Trade, Ontario, actual

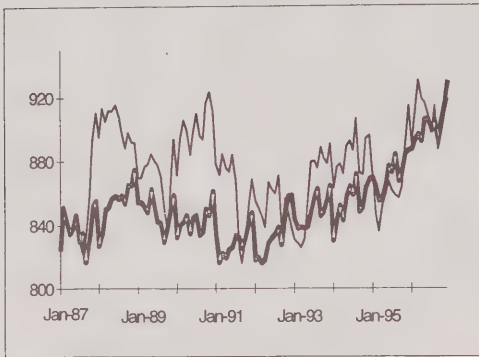
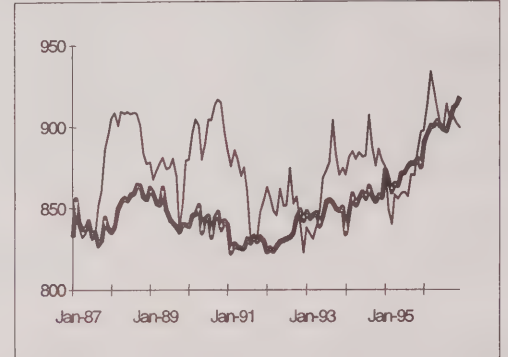


Figure 2(b) Employment in Trade, Ontario, seasonally adj.



$$C_{t(rca)} = F_t + b_{t(rca)}[(1-\alpha)(\tilde{C}_{t-1(rca)} - \bar{D}_{t-1}^* + \bar{B}_t - F_t) + \alpha(\tilde{C}_{t-1(rca)} - \bar{D}_{t-1}^*)] \quad (6.1)$$

where α is prescribed (1/3, say, but in general could be y-specific), and the coefficient $b_{t(rca)}$ is computed using the gr-system as in rc-class of estimates. A simple interpretation of (6.1) can be obtained by comparing with the ak^* -estimator of (3.7). First write (3.7) as

$$C_{t(ak^*)} = F_t + k^*[(1 - a^*/k^*)(\tilde{C}_{t-1(ak^*)} - \bar{D}_{t-1}^* + \bar{B}_t - F_t) + (a^*/k^*)(\tilde{C}_{t-1(ak^*)} - \bar{D}_{t-1}^*)]. \quad (6.2)$$

Now, for (6.1), α can be roughly viewed as the ratio of the two optimal coefficients a^*, k^* , and the factor k^* outside the square brackets of (6.2) is replaced by the

(suboptimal) regression coefficient $b_{t(rca)}$. Thus $C_{t(rca)}$ is not equivalent to the optimal ak^* -estimator, but some optimality could be preserved (if α is made y-specific) in setting the relative contribution of change and level driven predictors. Note, however, that the problem of internal inconsistency as mentioned in the introduction might arise if α is y-specific. Other attractive features of this version are that the value of α can be chosen to be well bounded away from zero (this should help to avoid sustained gaps between gr and rc series), and the number of extra controls is not doubled when both level and change driven predictors are included, thus allowing for introducing more controls as well as more degrees of freedom in variance estimation.

As a diagnostic of the impact of bias due to imputation of the previous month's employment status in view of the nonresponse of some of the present month respondents, the following simple check can be performed. The basic idea is to compute a multiplicative bias adjustment factor to the

estimator \bar{D}_{t-1}^* involving imputed values. The factor is defined as the ratio of two gr-estimators of the previous month's characteristic based on the matched subsample. The denominator is a gr-estimator for the previous month (involving imputed values) while the numerator is a gr-estimator for the previous month (not involving imputed values), both computed in a somewhat nonstandard way. For the numerator, we use the time $t-1$ respondents with their time $t-1$ responses, and after nonresponse adjustment of the design weights, construct the gr-estimator with controls for time t . For the denominator, we assume that the subsets of each of the matched subsamples at $t-1$ and t (here the matching is done with respect to each other, one forward in time and the other backward) not having the counterpart because of nonresponse, are statistically exchangeable with respect to each other. We then replace the time $t-1$ respondents who did not respond at time t by the time $t-1$ nonrespondents who responded at t , along with their imputed time $t-1$ responses as well as design weights. Now the nonresponse, and gr-poststratification (with controls for t) weight adjustments are redone for this modified full sample at $t-1$. The gr-weights so obtained are used to compute the denominator mentioned above. One can now look at the time series of this factor over several months for diagnostics on the bias due to imputation. If this is not deemed close to one, then the average of the factor over several months can be treated as a nonrandom multiplicative bias adjustment to \bar{D}_{t-1}^* . In practice, instead of adjusting \bar{D}_{t-1}^* , it would be preferable computationally to adjust the new control $\tilde{C}_{t-1(\text{rc})}$ (of equation 3.6) for the corresponding characteristic by inverse of the above multiplicative factor. Alternatively, the need for imputation can be avoided altogether if the questionnaire can be modified to obtain the necessary past information as suggested in the introduction.

The study of Lent, Miller and Cantwell (1994, 1996) considers the ak-composite weighted estimator for the U.S. Current Population Survey as an alternative to the currently used ak -estimator with $a=0.2$, $k=0.4$. Based on our experience with ak^* , it may be recommended that the ak^* -composite weighted estimator might be a better alternative in the interest of efficiency gains.

ACKNOWLEDGEMENTS

The bulk of this research work was done when the first and third authors were at Statistics Canada. The authors are indebted to M. Sheridan, J.D. Drew, J. Gambino and especially M.P. Singh for their encouragement and several useful discussions. They are grateful to Jon Rao and Wayne Fuller for comments and suggestions. They are also grateful to J.M. Levesque, P. Lorenz, and especially T. Merkouris (with whom this work initially got underway) for their assistance in analysis and interpretation of results. Thanks are also due to the referee and the assistant editor Harold

Mantel for their useful suggestions for revision of the paper. The first author's research was supported in part by an NSERC grant held at Carleton University under an adjunct research professorship.

REFERENCES

- BAILAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics*, 6: Sampling, Elsevier Science, NY, 187-211.
- CASSEL, C.M., SÄRNDAL, C.-E. and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd edition. John Wiley and Sons.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- FULLER, W.A. (1999). The Canadian Regression Composite Estimation. Unpublished manuscript.
- GODAMBE, V.P., and THOMPSON, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal Statistical Planning and Inference*, 22, 137-172.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.
- HANSEN, M.H., HURWITZ, W.N. and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, 2. John Wiley and Sons.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.
- LENT, J., MILLER, S. and CANTWELL, P. (1994). Composite weights for the current population survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 867-872.
- LENT, J., MILLER, S. and CANTWELL, P. (1996). Effects of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section Survey Research Methods, American Statistical Association*. 1, 130-139.
- LIANG, K.-Y., and ZEGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SÄRNDAL, C.-E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

- SINGH, A.C. (1994). Sampling-design-based estimating functions for finite population totals. Invited paper, *Abstracts of the Annual Meeting of the Statistical Society of Canada, Banff, Alberta, May 8-11*, p. 48.
- SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods, American Statistical Association*, 1, 120-129.
- SINGH, A.C., and FOLSOM, R.E. Jr. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods, American Statistical Association*, 610-615.
- SINGH, A.C., KENNEDY, B., WU, S. and BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.
- SINGH, A.C., and MERKOURIS, P. (1995). Composite Estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.
- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

A Regression Composite Estimator with Application to the Canadian Labour Force Survey

WAYNE A. FULLER and J.N.K. RAO¹

ABSTRACT

The Canadian Labour Force Survey is a monthly survey of households selected according to a stratified multistage design. The sample of households is divided into six panels (rotation groups). A panel remains in the sample for six consecutive months and is then dropped from the sample. In the past, a generalized regression estimator, based only on the current month's data, has been implemented with a regression weights program. In this paper, we study regression composite estimation procedures that make use of sample information from previous periods and that can be implemented with a regression weights program. Singh (1996) proposed a composite estimator, called MR2, which can be computed by adding x -variables to the current regression weights program. Singh's estimator is considerably more efficient than the generalized regression estimator for one-period change, but not for current level. Also, the estimator of level can deviate from that of the generalized regression estimator by a substantial amount and this deviation can persist over a long period. We propose a "compromise" estimator, using a regression weights program and the same number of x -variables as MR2, that is more efficient for both level and change than the generalized regression estimator based only on the current month data. The proposed estimator also addresses the drift problem and is applicable to other surveys that employ rotation sampling.

KEY WORDS: Survey sampling; Rotating samples; Combining estimators.

1. INTRODUCTION

Composite estimation is a term used in survey sampling to describe estimators for a current period that use information from previous periods of a periodic survey with a rotating design. When some units are observed in some of the periods, but not in all periods, it is possible to use this fact to improve estimates for all time periods.

Statistics Canada, U.S. Bureau of the Census and some other statistical agencies use a rotating design for labour force surveys. The current Canadian Labour Force Survey (LFS) is a monthly survey of about 59,000 households, which are selected according to a stratified multistage sampling design. The ultimate sampling unit is the household and a sample of households is divided into six panels (rotation groups). A rotation group remains in the sample for six consecutive months and is then dropped from the sample completely. Thus five-sixths of the sample of households is common between two consecutive months. Singh, Drew, Gambino and Mayda (1990) and Gambino, Singh, Dufour, Kennedy and Lindeyer (1998) contain detailed descriptions of the LFS design. In the U.S. Current Population Survey (CPS), the sample is composed of eight rotation groups. A rotation group stays in the sample for four consecutive months, leaves the sample for the succeeding eight months, and then returns for another four consecutive months. It is then dropped from the sample completely. Thus there is a 75 percent month-to-month sample overlap and a 50 percent year-to-year sample overlap (Hansen, Hurwitz, Nisselson and Steinberg 1955).

Patterson (1950), following the initial work by Jessen (1942), provided the theoretical foundations for design and

estimation for repeated surveys, using generalized least squares procedures. For the CPS, Hansen *et al.* (1955) proposed a simpler estimator, called the K -composite estimator. Gurney and Daly (1965) presented an improvement to the K -composite estimator, called the AK -composite estimator with two weighting factors A and K . Breau and Ernst (1983) compared alternative estimators to the K -composite estimator for the CPS. Rao and Graham (1964) studied optimal replacement schemes for the K -composite estimator. Eckler (1955) and Wolter (1979) studied two-level rotation schemes such as the one used in the U.S. Retail Trade Survey. Yansaneh and Fuller (1998) studied optimal recursive estimation for repeated surveys. Fuller (1990) and Lent, Miller, Cantwell and Duff (1999) developed the method of composite weights for the CPS. The composite weights are obtained by raking the design weights to specified control totals that included population totals of auxiliary variables and K -composite estimates for characteristics of interest, y . Using the composite weights, users can generate estimates from microdata files for the current month without recourse to data from previous months.

The above authors used the traditional design-based approach, assuming the unknown totals on each occasion to be fixed parameters. Other authors (Scott, Smith and Jones 1977; Jones 1980; Binder and Dick 1989; Bell and Hillmer 1990; Tiller 1989 and Pfeffermann 1991) developed estimates for repeated surveys under the assumption that the underlying true values constitute a realization of a time series.

Statistics Canada considered K and AK composite estimation for the Labour Force Survey at several times during the past 25 years (Kumar and Lee 1983), but did not

¹ Wayne A. Fuller, Iowa State University, 221 Snedecor Hall, Ames, IA, U.S.A.; J.N.K. Rao, Carleton University, Ottawa, Ontario K1S 5B6.

adopt composite estimation. Instead, a generalized regression estimator, based only on the current months data, has been computed with a regression weights program. When composite estimation was considered in the 1990's, there was strong pressure to develop a composite estimation procedure that used the existing estimation program. Singh (1996) proposed an ingenious method, called Modified Regression (MR), to address this issue. This method leads to a composite estimator, called MR2 estimator, which uses the existing regression weights program. Singh suggested creating x -variables to be used as control variables in the regression program. With the created variables and the previous period estimator, the existing regression weights program is used to construct regression weights that define the estimator for the current period. Control variables with known population totals are also included.

An empirical study of the MR2 estimator identified several characteristics of the procedure. First, the estimated variance of a one-period change is much reduced. Second, the estimated variance of level is often similar to that for the direct estimator. Third, the estimator of level could deviate from the direct estimator by a substantial amount and this deviation could extend over a long period.

In this paper, we study the efficiency of MR estimators theoretically, under a simplified set-up. We propose also a "compromise" estimator that leads to significant gains in efficiency, for both level and change, over the estimator using only the current month's data. The composite estimator also addresses the "drift" problem mentioned above and can be implemented using the existing regression weights program. Gambino, Kennedy and Singh (2000) evaluated the efficiency of the composite estimates for the LFS data, using a jackknife method of variance estimation. Bell (2000) compared several composite estimators using data from the Australian Labour Force Survey.

2. COMPOSITE REGRESSION ESTIMATION

There are two types of observations used in composite estimation; those observed only at the current time, t , and those observed both at the current time and at the previous time, $t-1$. Sometimes information in previous observations is condensed in the estimate(s) for the previous period(s). Let w_i be the sampling weight for observation i at time t , let A_t be the set of elements with observations at both the time periods t and $t-1$, and let B_t be the set of elements observed only at the current time period t . In this initial context, i is the index for an individual respondent. If there is no nonresponse, the set A_t for the LFS is composed of individuals in the five panels that were in the sample during the previous period, called the overlap panels. With no nonresponse, the set B_t for the LFS contains individuals first observed in the current period, called the birth panel. Assume

$$\sum_{i \in A_t} w_i + \sum_{i \in B_t} w_i = \hat{N}_t = \text{estimated population total.}$$

Let θ_t be the fraction of the sample in the overlap at time t :

$$\theta_t = \hat{N}_t^{-1} \sum_{i \in A_t} w_i. \quad (2.1)$$

In the Labour Force Survey θ_t is about 5/6 and is nearly constant over time. We will frequently omit the subscript t on A_t , B_t and θ_t , for simplicity.

2.1 Estimator

Singh's (1996) MR2 estimator uses the control variable

$$\begin{aligned} x_{1ti} &= \theta_t^{-1} (y_{t-1,i} - y_{ti}) + y_{ti} & \text{if } i \in A_t \\ &= y_{ti} & \text{if } i \in B_t, \end{aligned} \quad (2.2)$$

in the regression program, where y_{ti} is the value of a characteristics of interest, y , for element i at time t . Because of nonresponse in the LFS, Singh's original proposal used imputation for missing data and set $\theta = 5/6$, after imputation for missing data. In our initial discussion we use the θ_t as defined in (2.1), assuming no nonresponse so that imputation is not required. Note that "micromatching" of individual data files at $t-1$ and t is needed to calculate x_{1ti} and the resulting MR2 estimator. Additional control variables of the form (2.2) associated with other y -variables as well as auxiliary variables with known population totals are also included in the regression estimation. The auxiliary variables in the LFS include demographic variables such as age, sex and location.

The particular x -variables in (2.2) is designed such that the estimated total of x_1 is an estimator of the previous period total of y . Thus, the control total for x_1 in the regression procedure is the previous period estimator of the total of y .

Let $\hat{\mu}_{t-1}$ be the estimator of the mean of y for period $t-1$, let $\bar{y}_{m,t-1}$ and \bar{y}_{mt} be the means of the matched panels at time $t-1$ and t respectively, let \bar{y}_t be the grand mean of all sample panels at time t , and let \bar{y}_{B_t} be the mean of the birth panel at time t . Assume the sample of size n is divided into g panels of equal size and denote the matched sampling fraction by θ . To simplify the discussion we consider a single y -variable. Then Singh's (1996) MR2 estimator at time t , constructed with x_{1ti} , can be written in a regression estimator form as

$$\hat{\mu}_t = \bar{y}_t + (\bar{x}_{CNt} - \bar{x}_{Ct}) \hat{\beta}_{Ct} + [\hat{\mu}_{t-1} - (\bar{y}_{m,t-1} - \bar{y}_{mt} + \bar{y}_t)] b_t, \quad (2.3)$$

where \bar{x}_{CNt} is the population mean of the vector of auxiliary variables, such as age and sex, at time t , \bar{x}_{Ct} is the weighted sample mean of the auxiliary variables, and $(\hat{\beta}_{Ct}', b_t)'$ is the vector of regression coefficients for the regression of y_t on (x_{Ct}, x_{1t}) .

One can write

$$y_{t,i} = \hat{y}_{t,i(r)} + d_{t,i(r)},$$

where $\hat{y}_{t,i,(r)}$ is the predicted value in the regression of $y_{t,i}$ on x_{Cr} and $d_{t,i,(r)}$ is the deviation from the regression predicted value. Then

$$\begin{aligned} x_{1t} &= \theta^{-1} (\hat{y}_{t-1,i,(t-1)} + d_{t-1,i,(t-1)} - \hat{y}_{t,i,(t)} - d_{t,i,(t)}) \\ &\quad + \hat{y}_{t,i,(t)} + d_{t,i,(t)} \quad \text{if } i \in A_t \\ &= \hat{y}_{t,i,(t)} + d_{t,i,(t)} \quad \text{if } i \in B_t. \end{aligned}$$

For demographic variables X_{Cti} , it is reasonable to believe that $\hat{y}_{t-1,i,(t-1)}$ is close to $\hat{y}_{t,i,(t)}$ and close to $\hat{y}_{t,i,(t)}$. Therefore the part of x_{1t} that is orthogonal to x_{Cr} is close to

$$\begin{aligned} x_{d,1,t} &= \theta^{-1} (d_{t-1,i,(t-1)} - d_{t,i,(t)}) \\ &\quad + d_{t,i,(t)} \quad \text{if } i \in A_t \\ &= d_{t,i,(t)} \quad \text{if } i \in B_t. \end{aligned}$$

Thus the partial regression coefficient b_t is close to the regression coefficient for the regression of $d_{t,i,(t)}$ on $x_{d,1,t}$, and the value depends on the correlation between $d_{t,i,(t)}$ and $d_{t-1,i,(t-1)}$. A simple model for $d_{t,i,(t)}$ that has been used in the past, and the one we adopt in our analysis, is the assumption that the $d_{t,i,(t)}$ is the sum of a fixed μ_t and an error that is a first order autoregression with parameter ρ .

To simplify the presentation, we discuss the simple random sampling model without x_{Cr} . The results extend to the general case by considering the parameter ρ to be the partial correlation between y_t and y_{t-1} after adjusting for x_{Cr} .

Under the autoregressive model with fixed ρ , an intercept and no other x_{Cr} in the model, it can be shown that b_t converges in probability to

$$b_0 = \rho \lim_{n \rightarrow \infty} b_t = \theta \rho [2 - \theta - 2(1 - \theta)\rho - (1 - \theta)\sigma_y^{-2}\Delta_t^2]^{-1},$$

where $\Delta_t^2 = (\mu_t - \mu_{t-1})^2$. Assuming $(1 - \theta)\sigma_y^{-2}\Delta_t^2$ is small relative to the other terms we get

$$b_0 \doteq \theta \rho [2 - \theta - 2(1 - \theta)\rho]^{-1}. \quad (2.4)$$

For the LFS, $b_0 = (7 - 2\rho)^{-1}5\rho$.

Alternative representations for the estimator $\hat{\mu}_t$, omitting x_{Cr} , are obtained using the formula $\bar{y}_t = \theta \bar{y}_{mt} + (1 - \theta)\bar{y}_{Bt}$. Thus

$$\begin{aligned} &= (1 - b)\bar{y}_t + [\hat{\mu}_{t-1} + (\bar{y}_{mt} - \bar{y}_{m,t-1})]b \\ &= \theta [\bar{y}_{mt} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b] \\ &\quad + (1 - \theta)(\hat{\mu}_{t-1} - \bar{y}_{m,t-1}\bar{y}_{m,t-1} + \bar{y}_{mt})b + (1 - \theta)(1 - b)\bar{y}_t \\ &= [\theta + (1 - \theta)b][\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b] \\ &\quad + (1 - \theta)(1 - b)\bar{y}_{Bt} \\ &= \lambda_A [\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b^*] + (1 - \lambda_A)\bar{y}_{Bt}, \end{aligned} \quad (2.5)$$

where

$$1 - \lambda_A \approx (1 - \theta)(1 - b_0)$$

and

$$b^* \approx [\theta + (1 - \theta)b_0]^{-1}b_0. \quad (2.6)$$

The first expression on the right of the equality of (2.5) gives the MR2 estimator as a linear combination of the direct estimator \bar{y}_t and the difference estimator $\hat{\mu}_{t-1} + (\bar{y}_{mt} - \bar{y}_{m,t-1})$ i.e., in the form of a composite estimator. The final expression of (2.5) gives the estimator as a linear combination of a "regression-type" estimator based on the overlap panels and the mean of the birth panels.

2.2 An Alternative Estimator

It is possible to define alternative regression variables to use in regression composite estimation. We present a particular regression variable in this subsection. The associated regression estimator is not suggested as the ultimate estimator, but the estimator is a member of a class for which efficiency calculations are given. An alternative to Singh's (1996) MR2 estimator is outlined in section 5.

Define a variable to be equal to the previous period value if the individual is in the overlap sample and to be equal to the estimated mean for the previous period if the individual is in the birth sample. The regression variable is

$$\begin{aligned} x_{2,ti} &= y_{t-1,i} \quad \text{if } i \in A_t \\ &= \hat{\mu}_{t-1} \quad \text{if } i \in B_t. \end{aligned} \quad (2.7)$$

If this variable is used in a regression estimator, the control mean is $\hat{\mu}_{t-1}$, the previous period estimator, because the mean for the created variable is estimating the mean for period $t - 1$. Singh (1996) used a variable $\tilde{x}_{2,ti}$ similar to $x_{2,ti}$. In Singh's variable, the $\hat{\mu}_{t-1}$ in (2.7) is $\bar{y}_{m,t-1}$ if $i \in B_t$.

Consider a regression estimator constructed with $x_{2,ti}$ and recall that the control mean of $x_{2,t}$ is $\hat{\mu}_{t-1}$. The regression estimator using $x_{2,t}$ can be written

$$\hat{\mu}_{reg,t} = \bar{y}_t + (\hat{\mu}_{t-1} - \bar{x}_{2,t})\hat{\beta}, \quad (2.8)$$

where $\hat{\beta}$ is the regression coefficient for the regression of y_t on $x_{2,t}$ (subscript t is dropped on $\hat{\beta}_t$ for simplicity), \bar{y}_t is the sample mean of y at time t , and $\bar{x}_{2,t}$ is the sample mean of $x_{2,ti}$ for all sample panels at time $t - 1$. The regression coefficient $\hat{\beta}$ is, approximately, the regression of y_t on $x_{2,t}$ in the set A . The coefficient is not exactly the regression coefficient for the set A because $\bar{y}_{m,t-1}$ is not equal to $\hat{\mu}_{t-1}$, but the difference between the two estimators will usually be small. Singh (1996) called the regression estimator constructed with $\tilde{x}_{2,ti}$, the MR1 estimator.

Using $\bar{y}_t = \theta \bar{y}_{mt} + (1 - \theta)\bar{y}_{Bt}$, the regression estimator of μ_t using $x_{2,t}$ as a control variable is given by

$$\hat{\mu}_t = (1 - \theta)\bar{y}_{B,t} + \theta \{ \bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})\hat{\beta} \}. \quad (2.9)$$

The expression within curly brackets in (2.9) is the regression estimator of μ_t using the estimator $\hat{\mu}_{t-1}$ and only the data from the matched sample A . Note that the regression estimator

$$\hat{\mu}_{m,t} = \bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})\beta, \quad (2.10)$$

where β is the regression of y_t on y_{t-1} in the set A , is the optimal linear estimator for μ_t based on $\hat{\mu}_{t-1}$ and the data of set A . Note that $\beta = \rho$ if the variances are the same at the two time periods. Hereafter, we often set $\beta = \rho$.

Using the variable x_{2t} gives the optimal estimator, $\hat{\mu}_{mt}$, based on data in set A , but it does not combine that estimator with the mean of set B in an optimal way. As can be seen in (2.10), the weight given to the mean of set B is $1 - \theta$. In general, this weight is too large because the variance of the regression estimator is less than the variance of the simple mean.

3. OPTIMAL ESTIMATION

The way in which one chooses to combine the regression estimator for set A with the mean of set B depends on one's objective function and on the variance of $\hat{\mu}_{t-1}$. We give some illustrative calculations based on some simplifying assumptions. For convenience let $V\{\hat{\mu}_{t-1}\}$ be expressed as a multiple of the variance of the birth panel,

$$V\{\hat{\mu}_{t-1}\} = q_t^{-1} V\{\bar{y}_{B,t}\}. \quad (3.1)$$

Assume

$$V\{\bar{y}_t\} = g^{-1} V\{\bar{y}_{B,t}\}, \quad (3.2)$$

$$\text{Cov}\{\hat{\mu}_{t-1}, (\bar{y}_{m,t} - \bar{y}_{m,t-1})\beta\} = 0, \quad (3.3)$$

$$\text{Cov}\{\hat{\mu}_{t-1}, \bar{y}_{B,t}\} = 0, \quad (3.4)$$

and

$$\text{Cov}\{\bar{y}_{B,t}, (\bar{y}_{m,t} - \bar{y}_{m,t-1})\beta\} = 0, \quad (3.5)$$

where g is the number of rotation groups (panels). Assumption (3.1) is reasonable if the original panels have a covariance function well approximated by that of a first order autoregressive process. For the LFS, the zero covariances in (3.4) and (3.5), and assumption (3.2) are only approximations because $\bar{y}_{B,t}$ is not based on an entirely independent sample.

We write the regression estimator based on the overlap as

$$\hat{\mu}_{m,t} = \bar{y}_{m,t} - \bar{y}_{m,t-1}\beta + \hat{\mu}_{t-1}\beta$$

and, with the assumptions, obtain

$$V(\hat{\mu}_{mt}) = [g^{-1}\theta^{-1}(1 - \rho^2) + q_t^{-1}\rho^2] V\{\bar{y}_{B,t}\}. \quad (3.6)$$

For the LFS, $g = 6$ is the number of panels. Now consider an estimator that is a linear combination of $\hat{\mu}_{m,t}$ and $\bar{y}_{B,t}$,

$$\begin{aligned} \hat{\mu}_t &= \lambda \hat{\mu}_{mt} + (1 - \lambda) \bar{y}_{B,t} \\ &= \lambda (\bar{y}_{m,t} - \bar{y}_{m,t-1}\beta + \hat{\mu}_{t-1}\beta) + (1 - \lambda) \bar{y}_{B,t}, \end{aligned} \quad (3.7)$$

where $0 \leq \lambda \leq 1$ is to be determined. To minimize the variance of current level, given $\hat{\mu}_{t-1}$ with variance $q_t^{-1} V\{\bar{y}_{B,t}\}$, one would minimize

$$\begin{aligned} V\{\hat{\mu}_t\} &= V\{\lambda \hat{\mu}_{mt} + (1 - \lambda) \bar{y}_{B,t}\} \\ &= \lambda^2 V\{\hat{\mu}_{mt}\} + (1 - \lambda)^2 V\{\bar{y}_{B,t}\}, \end{aligned} \quad (3.8)$$

with respect to λ . Under the assumptions (3.3), (3.4) and (3.5), the optimum λ for current level is

$$\lambda_{\text{opt}} = [g^{-1}\theta^{-1}(1 - \rho^2) + q_t^{-1}\rho^2 + 1]^{-1}.$$

However, if one is planning on using the estimator for a long period of time, one must realize that only certain values of q_t are possible in the long run. The value of λ chosen to estimate μ_t determines the variance of $\hat{\mu}_t$ and hence, determines the variance that will go into the estimator of μ_{t+1} . Assuming $\beta = \rho$, we have

$$V\{\hat{\mu}_t\} = \{g^{-1}\theta^{-1}\lambda^2(1 - \rho^2) + q_t^{-1}\lambda^2\rho^2 + (1 - \lambda)^2\} V\{\bar{y}_{B,t}\}$$

or

$$q_{t+1}^{-1} = g^{-1}\theta^{-1}\lambda^2(1 - \rho^2) + (1 - \lambda)^2 + \lambda^2\rho^2 q_t^{-1}. \quad (3.9)$$

Thus, for a given λ , the limiting value for q_t^{-1} is

$$\begin{aligned} \lim_{t \rightarrow \infty} q_t^{-1} &= (1 - \lambda^2\rho^2)^{-1} [g^{-1}\theta^{-1}\lambda^2(1 - \rho^2) \\ &\quad + (1 - \lambda)^2]. \end{aligned} \quad (3.10)$$

This result is equivalent to that given by Cochran (1977), page 352 equation (12.86).

Table 1 contains values of the limit variances as the number of periods becomes large, for selected values of ρ and λ , where $\theta = 5/6$ and $g\theta = 5$ for the LFS. The variances are standardized so that the variance of the direct estimator based on the mean of six panels is 1.00. Thus, the entries are six times the limiting value in (3.10). If the correlation is 0.95 and λ is set equal to 0.96, the long run variance of current level is 70 % of that of the direct estimator. If λ is set equal to 0.90, the long run variance is 58 % of that of the direct estimator when $\rho = 0.95$.

The first line in Table 1 is for $\lambda = 5/6 = \theta$. This is the λ corresponding to the use of x_{2t} in a regression estimator. The variance with $\lambda = 5/6$ is always smaller than that of the direct estimator because of the improvement associated

with the use of the regression estimator $\hat{\mu}_{m,t}$. Thus, if $\rho \neq 0$, the regression estimator with $x_{2,t}$ leads to significant reduction in variance over the direct estimator, \bar{y}_t , that uses current data only.

Table 1
Standardized Limit Variances of Level:
LFS Rotation Pattern

λ	ρ				
	0.70	0.80	0.90	0.95	0.98
0.833	0.897	0.840	0.743	0.665	0.600
0.840	0.895	0.836	0.734	0.650	0.581
0.860	0.894	0.830	0.714	0.614	0.527
0.880	0.903	0.835	0.705	0.588	0.481
0.900	0.921	0.851	0.711	0.575	0.444
0.920	0.951	0.882	0.736	0.582	0.420
0.940	0.992	0.928	0.785	0.617	0.420
0.960	1.046	0.994	0.867	0.698	0.465
0.980	1.115	1.083	0.997	0.861	0.619
0.990	1.155	1.138	1.087	0.998	0.803
0.995	1.177	1.168	1.140	1.089	0.960

The optimal λ is a function of ρ and increases slowly as ρ increases. For $\rho = 0.0$, the optimal λ is 0.833, for $\rho = 0.7$ the optimal λ is about 0.85, for $\rho = 0.95$ the optimal λ is about 0.91 and for $\rho = 0.98$ the optimal λ is about 0.93.

We now turn to the MR2 estimator (2.3) which can be written as

$$\hat{\mu}_t = \lambda_A [\bar{y}_{m,t} + (\hat{\mu}_{t-1} - \bar{y}_{m,t-1})b^*] + (1 - \lambda_A)\bar{y}_{B,t},$$

where λ_A and b^* are defined in (2.6). While the MR2 estimator is not a member of the class (3.7), to the degree that b^* is “close to” ρ , it is “close to” a member of the class. For example if $\rho = 0.95$, then $b_0 \doteq 0.9314$ and $b^* \doteq 0.9422$. If $\rho = 0.90$, then $b_0 \doteq 0.8659$ and $b^* \doteq 0.8853$.

Using the limiting value b_0 of b , we have $(1 - \lambda_A) = (1 - \theta)(1 - b_0)$, where b_0 is given by (2.4). Then $\lambda_A = 0.9375, 0.9568, 0.9776, 0.9886$, and 0.9954 for $\rho = 0.70, 0.80, 0.90, 0.95$ and 0.98 , respectively. From Table 1, the standardized variances of $\hat{\mu}_t$ for these values of λ_A are $0.986, 0.982, 0.978, 0.976$, and 0.975 , for $\rho = 0.70, 0.80, 0.90, 0.95$, and 0.98 , respectively. Thus, the MR2 estimator for current level has an efficiency for level that is essentially the same as that of the direct estimator, \bar{y}_t . The efficiency of the MR1 estimator is that for $\lambda = 0.833$ in Table 1 and is always superior to that of \bar{y}_t .

4. VARIANCE OF ONE-PERIOD CHANGE

Given $\hat{\mu}_{t-1}$, $\bar{y}_{m,t-1}$, $\bar{y}_{m,t}$ and $\bar{y}_{B,t}$ the optimal estimator of μ_{t-1} is no longer $\hat{\mu}_{t-1}$ because $\bar{y}_{m,t}$ contains information about μ_{t-1} . However, it is not customary practice to revise the estimator of μ_{t-1} . Given no revision, the estimator of change is $\hat{\mu}_t - \hat{\mu}_{t-1}$.

Under no revision in $\hat{\mu}_{t-1}$ and conditions (3.2) through (3.5), the variance of $\hat{\mu}_t - \hat{\mu}_{t-1}$, where $\hat{\mu}_t$ is defined in (3.7), is

$$\begin{aligned} V\{\hat{\mu}_t - \hat{\mu}_{t-1}\} &= V\{\lambda[\bar{y}_t + (\hat{\mu}_{t-1} - \bar{x}_{2,t})\rho] \\ &\quad + (1 - \lambda)\bar{y}_{B,t} - \hat{\mu}_{t-1}\} \\ &= [g^{-1}\theta^{-1}\lambda^2(1 - \rho^2) + (1 - \lambda)^2 \\ &\quad + (\rho\lambda - 1)^2 q_t^{-1}] V\{\bar{y}_{B,t}\}. \end{aligned} \tag{4.1}$$

Table 2 contains standardized limit variances of the estimated change, $\hat{\mu}_t - \hat{\mu}_{t-1}$, for selected values of g and λ , with $g\theta = 5$. The entries in the table are the limiting variances of estimated change divided by the variance of change based on the common elements, $V\{\bar{y}_{m,t} - \bar{y}_{m,t-1}\}$. The variance of change based on the common elements is $2\theta^{-1}(1 - \theta)(1 - \rho)V\{\bar{y}_{B,t}\}$.

Table 2
Standardized Limit Variances of No-Revision
One Period Change: LFS Rotation Pattern

λ	ρ				
	0.70	0.80	0.90	0.95	0.98
0.833	1.039	1.168	1.550	2.312	4.595
0.840	1.024	1.142	1.492	2.189	4.277
0.860	0.989	1.079	1.345	1.872	3.454
0.880	0.963	1.029	1.223	1.607	2.756
0.900	0.947	0.993	1.127	1.391	2.181
0.920	0.940	0.970	1.055	1.222	1.723
0.940	0.942	0.959	1.007	1.100	1.379
0.960	0.953	0.961	0.982	1.024	1.146
0.980	0.972	0.975	0.980	0.991	1.021
0.990	0.985	0.986	0.987	0.990	0.998
0.995	0.992	0.993	0.993	0.994	0.996

Tables 1 and 2 make clear the cost of not revising the estimate of $\hat{\mu}_{t-1}$. For example, if $\rho = 0.95$, the variance of no-revision one period change is minimized with $\lambda \doteq 0.99$, but the variance of level is minimized with $\lambda \doteq 0.91$. A compromise value of $\lambda = 0.95$ gives a variance of level that is about 14 % larger than optimal and a variance of change that is about 7 % larger than the smallest variance of Table 2.

Using the values of λ_A associated with the MR2 estimator, the entries in Table 2 are 0.940, 0.960, 0.979, 0.989, and 0.996 for $\rho = 0.70, 0.80, 0.90, 0.95$ and 0.98 , respectively. Thus, given no revision, and ignoring the difference between b_0 and ρ , the MR2 estimator is nearly optimal as an estimator of change, unlike the MR1 estimator, where the MR1 estimator corresponds to $\lambda = 0.833$ in Table 2.

5. A COMPROMISE ESTIMATOR

On the basis of Table 2, the efficiency of the MR2 estimator of change for the LFS based on x_{1t} , for the no-revision case, is quite good. The MR1 no-revision estimator of change based on x_{2t} has relatively poor efficiency because it is a member of the class (3.7) with $\lambda = \theta = 0.8333$. On the other hand, the MR1 estimator of level based on x_{2t} is superior to the MR2 estimator based on x_{1t} , and there are members of the class (3.7) that are much superior to the MR2 estimator of level.

Because the λ in the MR2 estimator is relatively large and the λ for the MR1 estimator is relatively small, we can create approximations to most interesting members of the class (3.7) as linear combinations of (2.10) and (2.5). Let

$$x_{3,ti} = \alpha x_{1,ti} + (1 - \alpha)x_{2,ti}, \quad (5.1)$$

where $0 \leq \alpha \leq 1$ is a fixed number. The regression estimator based on $x_{3,ti}$ gives an approximation to a member of the class (3.7) with

$$\lambda = \alpha \lambda_A + (1 - \alpha)\theta, \quad (5.2)$$

where λ_A is defined in (2.6). Thus, if $\rho = 0.95$,

$$\lambda = \alpha (0.9886) + (1 - \alpha)(5/6),$$

for the LFS rotation pattern with $\theta = 5/6$ and $b_0 = (7 - 2\rho)^{-1}5\rho$; $\lambda = 0.95$ if $\alpha = 0.75$.

We choose α to give the desired combination of $\bar{y}_{B,t}$ and the "regression estimator" based on observations in set A. If one does not revise the estimator of μ_{t-1} , the preferred combination depends on the relative importance assigned to the variance of level and to the variance of change.

Table 3 gives the variance of the MR2 estimator ($\alpha = 1$) relative to the variance of the estimator constructed using $\alpha = 0.75$ and the variance of the estimator constructed using $\alpha = 0.65$. An entry in Table 3 for $\hat{\mu}_t$ is expression (3.10) evaluated at λ_A of (2.6) and ρ , divided by (3.10) evaluated at λ of (5.2) and ρ . An entry for $\hat{\mu}_t - \hat{\mu}_{t-1}$ is expression (4.1) evaluated at λ_A of (2.6) and ρ , divided by (4.1) evaluated at λ of (5.2) and ρ . These are approximations to actual efficiencies because ρ is used for the coefficient of x_3 . It is clear from Table 3 that the compromise estimator is slightly inferior to the MR2 estimator for one-period change, but is much superior to the MR2 estimator for level. For example, with $\rho = 0.95$ and $\alpha = 0.65$, the relative efficiency of the compromise estimator is 1.62 for level and 0.87 for one-period change.

For larger values of ρ , the variance of change is much smaller than the variance of level. Thus, for $\rho = 0.95$, the variance of level and of change for $\alpha = 1.00$ are about 1.00 and 0.12, respectively, while the variance of level and of change for $\alpha = 0.75$ are about 0.67 and 0.13, respectively, when expressed in common units.

The smaller α has the advantage that the composite estimator will be closer to the direct estimator. Thus,

potential biases associated with the composite estimator should be smaller with the smaller α .

Table 3
Approximate Efficiencies of Compromise
Estimators Relative to $\alpha = 1$

ρ	b_0	$\alpha = 0.75$			$\alpha = 0.65$	
		$1 - \lambda_A$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$
0.70	0.625	0.0625	1.052	0.999	1.069	0.995
0.80	0.741	0.0432	1.099	0.994	1.129	0.984
0.90	0.865	0.0224	1.238	0.975	1.303	0.946
0.95	0.931	0.0114	1.502	0.936	1.616	0.875
0.98	0.972	0.0046	2.177	0.833	2.321	0.712

6. DRIFT PROBLEM

As noted in the Introduction, the MR2 estimator could deviate from the direct estimator by a substantial amount and this deviation could extend over a long period. We now illustrate the basis for this phenomenon. We can express the deviation of the compromise regression estimator $\hat{\mu}_t$, based on $x_{3,ti}$, from the true mean μ_t as

$$\hat{\mu}_t - \mu_t = (\lambda\rho)^t (\hat{\mu}_0 - \mu_0) + \sum_{j=0}^{t-1} (\lambda\rho)^j [\lambda \bar{r}_{m,t-j} + (1 - \lambda)(\bar{y}_{B,t-j} - \mu_{t-j})], \quad (6.1)$$

where μ_0 is the mean at the initiation of the process and

$$\bar{r}_{mt} = \bar{y}_{mt} - \mu_t - \rho(\bar{y}_{m,t-1} - \mu_{t-1}).$$

If ρ is close to one and we use $\lambda = 1$, then the error $\hat{\mu}_t - \mu_t$ behaves roughly as a random walk which can lead to long periods in which $\hat{\mu}_t - \mu_t$ has the same sign. On the other hand, if $\alpha < 1$ and $\rho = 1$, then $\lambda < 1$ and the error $\hat{\mu}_t - \mu_t$ exhibits less drift. For example, if $\alpha = 0.70$, the correlation between adjacent errors $\hat{\mu}_t - \mu_t$, will be no greater than 0.95 under assumption (3.2)–(3.5). For the MR2 estimator, $\lambda \rightarrow 1$ as $\rho \rightarrow 1$ and hence the MR2 estimator can exhibit drift for ρ close to one.

7. CONCLUDING REMARKS

For simplicity, we often assumed simple random sampling to obtain theoretical results. Similar results hold for complex designs and additional auxiliary variables, by considering ρ to be a partial autocorrelation. Also, we used x_3 -variables corresponding to only one variable y , but several y -variables can be used in constructing the corresponding x -variables for use in regression estimation. Gambino, Kennedy and Singh (2001) conducted an empirical study with LFS data using several x_3 -variables with a common α , and arrived at a compromise α for use in the LFS.

In section 2.1, we assumed no nonresponse so that imputation is not required. But in the LFS, nonresponse on an item y may occur either at time $t - 1$ or a time t or at both time points. Gambino, Kennedy and Singh (2001) provide details of the imputation methods actually used in the LFS.

ACKNOWLEDGEMENTS

The research of Wayne Fuller was partly supported by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of the Census. We thank Harold Mantel for a careful reading of the manuscript that led to improvements.

REFERENCES

- BELL, P. (2001). Comparisons of alternative Labour Force Survey estimators. *Survey Methodology*, 27, 53-63.
- BELL, W.R. and HILLMER, S.C. (1990). The time series approach to estimation for periodic surveys. *Survey Methodology*, 16, 195-215.
- BINDER, D.A., and DICK, J.P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BREAU, P., and ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3rd Ed. New York : John Wiley and Sons.
- ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-180.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- GAMBINO, J.G., KENNEDY, B. and SINGH, M.P. (2001). Regression composite estimation for the Canadian labour force survey : Evaluation and implementation. *Survey Methodology*, 27, 65-74.
- GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. and LINDEYER, J. (1998). *Methodology of the Canadian Labour Force Survey*. Catalogue no. 71-526, Statistics Canada.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.
- HANSEN, M.H., HURWITZ, W.N., NISSELSON H. and STEINBERG, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.
- JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, B*, 42, 221-226.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 1-24.
- LENT, J., MILLER, S.M., CANTWELL, P.J. and DUFF, M. (1999). Effect of composite weights on some estimates from the Current Population Survey. *Journal of Official Statistics*, 14, 431-448.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, B*, 12, 241-255.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SCOTT, A.J., SMITH, T.M.F. and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.
- SINGH, M.P., DREW, J.D., GAMBINO, J. and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*. Catalogue No. 71-526, Statistics Canada.
- TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.
- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- YANSANEH, I.S., and FULLER, W.A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*, 24, 31-40.

Comparison of Alternative Labour Force Survey Estimators

PHILIP BELL¹

ABSTRACT

This paper looks at a range of estimators applicable to a regularly repeated household survey with controlled overlap between successive surveys. The paper shows how the Best Linear Unbiased Estimator (BLUE) based on a fixed window of time points can be improved by applying the technique of generalised regression. This improved estimator is compared to the *AK* estimator of Gurney and Daly (1965) and the modified regression estimator of Singh, Kennedy, Wu and Brisebois (1997), using data from the Australian Labour Force Survey.

KEY WORDS: Composite estimator; Best linear unbiased estimator; Modified regression; Repeated Surveys.

1. INTRODUCTION

This paper looks at a range of estimators applicable to a regularly repeated household survey with controlled overlap between successive surveys. The common theme of the estimators is to use data from previous times to improve current estimates, by taking advantage of correlations in the overlapping sample. I refer to all such estimators as composite estimators.

The estimators are evaluated for use in the Australian Labour Force Survey (LFS). In the LFS, overlap is controlled by dividing the first-stage sample of geographic areas into eight "rotation groups" from which dwellings are selected. In each month the same dwellings are selected from seven of the rotation groups, while new dwellings are selected in the remaining group. The sample consists of civilian persons aged 15 years old and over residing in the selected dwellings.

This sample design leads to high overlap of sample between two successive months within the seven "matched rotation groups". Using only data from these rotation groups rather than the whole sample can decrease the sampling error on an estimate of month to month movement. Composite estimation techniques seek to exploit this to give estimates with lower sampling error.

Section 2 of the paper introduces the Australian LFS and its current "generalised regression" estimator. The issue of time-in-survey bias (called rotation group bias by Bailer 1975) is also discussed.

Section 3 presents the "*AK* composite" estimator proposed by Gurney and Daly (1965). This method has been used in the US Current Population Survey for many years. An extension known as "*AK* composite weighting" has been used for the last few years; this was proposed by Fuller (1990) and studied by Lent, Miller and Cantwell (1994, 1996).

Section 4 presents the "modified regression" method of composite estimation (Singh and Merkouris 1995, Singh 1996). Here I focus on the MR2 estimator of Singh, *et al.*

(1997), which provides the largest reductions in sampling error. I also present a variant of this method suggested by Fuller (1999) for use in the Canadian Labour Force Survey.

Section 5 presents a "Best Linear Unbiased Estimator" (BLUE) based on data from a "window" containing a fixed number of successive months. This estimator was originally given by Jessen (1942) in the case of 2 occasions. A BLUE based on all occasions in a long series appears impractical, though a recursive approximation to this was developed by Yansaneh and Fuller (1998). This paper improves the fixed window BLUE described in Bell (1998) using the technique of generalised regression.

Section 6 gives the results of applying the different methods to the estimation of employed persons and unemployed persons in the LFS. Standard errors are estimated for longer-term indicators such as trend and trend movement, as well as for estimates of monthly level and its movement. Possible biases are explored, as well as evidence of change to seasonal patterns.

I conclude by comparing the advantages and disadvantages of the different types of estimator for application in the LFS. The improved BLUE estimator is found to be efficient, and when applied to the LFS is not subject to any large bias.

2. CURRENT ESTIMATES FOR THE LABOUR FORCE SURVEY

2.1 Overview of the LFS

The LFS has a multistage sample design, the first stage being a sample of small geographic areas known as "Census collector's districts" (CDs). A new sample of CDs is selected every five years, and the CDs are classified to eight "rotation groups". The dwellings selected from a CD remain in the sample for eight surveys, and then are replaced by other dwellings from the same CD. This replacement of dwellings is known as rotation, with all the dwellings in a rotation group being replaced at the same

¹ Philip Bell, Australian Bureau of Statistics, e-mail: philip.bell@abs.gov.au

time. Interviewers seek to collect data for all in-scope persons in the selected dwellings.

Of particular interest in the LFS is the person's labour force status – whether they are employed, unemployed or not in the labour force. The number of persons in each labour force status, for various categories of person, are key items to be estimated in the survey. Even more important to many users of the survey data than these level estimates are the estimates of movement in the figures between successive time points. It can be argued that longer-term indications of the direction of the series are even more important *e.g.*, the movement of the X11 trend or of a similar smoother (Bell 1999).

The sample design ensures that the unconditional probability of selection $\pi_{t,i}$ is known for each sampled person i at time t . This allows a simple estimator for a population total due to Horvitz and Thompson (1952). If Y_t is the population item to be estimated at time t , and y_{ti} is the same item as reported by the i -th unit at time t , the Horvitz-Thompson estimator is

$$\hat{y}_t^H = \sum_i w_{ti}^\pi y_{ti} \quad (1)$$

for $w_{ti}^\pi = \pi_{ti}^{-1}$, known as the selection weights.

2.2 The Generalised Regression (GR) Estimator

Generalised regression is a method for adjusting or “calibrating” a set of unit weights to add to a set of population attributes known as benchmarks. For a suitable choice of benchmarks the resulting weights give an improved estimate by taking account of externally available information.

In the LFS we start with the Horvitz-Thompson weights and calibrate them to add to demographic benchmarks that give numbers of people in the population for 560 poststrata (14 geographic regions classified by sex and 20 age groups). The weights from a given post-stratum are prorated to add to the stratum benchmark. This post-stratified ratio estimator is a particular case of the generalised regression or GR estimator.

Let x_{ti} be a row vector of auxiliary variables for unit i at time t , and $\hat{x}_t = \sum_i b_{ti} x_{ti}$ be estimates for the corresponding row vector of benchmark values X_t , based on some initial weights b_{ti} . The GR estimator based on these initial weights is then given by

$$\hat{y}_t^G = \hat{y}_t + (X_t - \hat{x}_t) \hat{\beta} \quad (2)$$

$$\text{for } \hat{\beta} = \left(\sum_i b_{ti} x_{ti}' x_{ti} \right)^{-1} \sum_i b_{ti} x_{ti}' y_{ti} \quad (3)$$

$$\text{i.e., } \hat{y}_t^G = \sum_i w_{ti}^G y_{ti} \text{ for}$$

$$w_{ti}^G = b_{ti} \left(1 + (X_t - \hat{x}_t) \left(\sum_i b_{ti} x_{ti}' x_{ti} \right)^{-1} x_{ti}' \right). \quad (4)$$

In post-stratified ratio estimation the row vectors x_{ti} contain zeroes except in the column corresponding to the unit's post-stratum, and b_{ti} are the selection weights w_{ti}^π . In this case the regression parameters are just the post-stratum means, estimated using the selection weights.

2.3 Rotation Group Estimates

Each rotation group consists of a representative sample of dwellings, and so can provide a separate estimate. Number the rotation groups at a time point according to the number of times the dwellings in the rotation group have been sampled. Write $R(t, i) = r$ if unit i is in the rotation group sampled for the r -th time at time t . The Horvitz-Thompson estimate of Y_t based on rotation group r is

$$\hat{y}_t^{Hr} = \sum_{i: R(t, i) = r} 8 w_{ti}^\pi y_{ti}. \quad (5)$$

Generalised regression can be used to improve these estimators, by calibrating the weights to add to a set of benchmarks. Unfortunately the lower sample size in a single rotation group may require using a smaller number of benchmarks than in the overall case. In the LFS situation I used a single generalised regression step on the whole sample so that across the whole sample the weights add to the benchmarks for the current 560 poststrata, while in each rotation group the weights add to an eighth of the benchmarks for 71 collapsed poststrata. The resulting weights, when applied to a given rotation group r and multiplied by eight, give the rotation group estimates \hat{y}_t^{Rr} .

2.4 Time-in-Survey Bias

Ideally rotation group estimates should have the same expectation Y_t , but in practice they have slightly different expectations, and hence different biases. Some of the difference is due to collection practices – for example, dwellings sampled for the first time are interviewed using a personal visit, while other rotation groups are mostly interviewed by telephone. It is not clear which rotation group is least affected by this sort of “time-in-survey” bias. The overall estimate will have a time-in-survey bias that is some mix of the biases from each rotation group. We rely on good survey methods to keep this bias small. Note that all the composite estimators will have different contributions from the rotation groups, and therefore different time-in-survey biases.

3. AK COMPOSITE ESTIMATION

3.1 AK Composite Estimator

The AK composite estimator (Gurney and Daly 1965) is designed to put extra emphasis on the movement from the matched rotation groups (those rotation groups in which the same dwellings were selected in the current and previous months). The estimator has three components. The first is a mean of the rotation group estimates for the current month

data (time t). The second is last month's AK composite plus a movement estimate based only on the matched rotation groups. The third component is the difference between estimates from the unmatched rotation group and from the matched ones. How much of each component to use is given by two parameters A and K , as follows:

$$\begin{aligned} \hat{y}_t^{AK} = & (1 - K) \frac{1}{8} \sum_{r=1}^8 \hat{y}_t^{Rr} \\ & + K \left(\hat{y}_{t-1}^{AK} + \frac{1}{7} \sum_{r=2}^8 \hat{y}_t^{Rr} - \frac{1}{7} \sum_{r=1}^7 \hat{y}_{t-1}^{Rr} \right) \\ & + A \left(\hat{y}_t^{R1} - \frac{1}{7} \sum_{r=2}^8 \hat{y}_t^{Rr} \right). \end{aligned} \quad (6)$$

3.2 Choosing Parameter Values

The key parameter is K , which gives how much of the new estimate is based on the matched rotation group movement. The optimal A and K to use will depend on the variable being estimated. Higher K values are appropriate for employment than for unemployment, since employment has a higher correlation between months.

AK composite estimates of persons employed, unemployed and "not in the labour force" will not add correctly to the total population unless the same parameters are used for all the estimates. This leads to using a compromise choice of A and K . The results in this report are based on $A = 0.06$ and $K = 0.7$. These values were found by trying a range of values of A and K , and choosing those that gave optimal employed estimates. In this study no values of A and K gave unemployed estimates appreciably better than these values.

Our empirical study did not show particularly good sampling errors for the AK estimator. The fine calibration that was used in obtaining the rotation group estimates may be to blame – it is possible that using broader categories would improve the sampling errors.

3.3 Properties of the AK Estimator

The AK estimator puts extra emphasis on the movement in the matched rotation groups. Thus the rotation group containing dwellings in sample for the first time contributes less than in the GR estimator. The AK estimator thus has a different time-in-survey bias to the GR estimator.

The AK estimator is recursive, in that last month's estimator is required in order to produce this month's. This is inconvenient for producing estimates for a new item or category. Also, the need to use the same values of A and K for all items can give sub-optimal estimates for any given item.

These concerns have led to the US Current Population Survey changing to a variant known as "AK composite weighting" (Lent, Miller and Cantwell 1994). In AK composite weighting, separate employed and unemployed estimates are produced for a number of important published

categories, using the AK composite with optimal parameters for the estimate in question. The current data is then calibrated so that the unit weights add to these AK estimates as well as demographic benchmarks. All estimates are then produced from the current dataset using these new "AK composite weights".

The convenience of producing all estimates as a weighted sum of a single month's data is a major advantage of the AK composite weighting approach. Another is that the most important estimates are AK composite estimates with near-optimal choice of AK . A disadvantage is that only the most important estimates are true composite estimates. Any other estimates (including estimates of persons not in the labour force) are typically not much improved over the standard GR estimates (Lent, Miller and Cantwell 1996).

4. MODIFIED REGRESSION ESTIMATION

4.1 Overview of Modified Regression

The modified regression method is another way to provide composite estimates that can be obtained as weighted aggregates of the current survey dataset. The method targets a predetermined set of key items, for which it achieves particularly low sampling errors.

The modified regression technique uses generalised regression on the current month's dataset after attaching new auxiliary variables z_{it} to each unit i at time t . Here z_{it} is a row vector with an element for each of the key items. Corresponding to these we have "pseudo-benchmarks" Z_t based on the previous month's estimates for the key items. The modified regression estimator is then given by a generalised regression step applying both the demographic benchmarks and the pseudo-benchmarks.

$$\hat{y}_t^M = \hat{y}_t^H + \left((X_t, Z_t) - (\hat{x}_t^H, \hat{z}_t^H) \right) \beta_t^M \quad (7)$$

$$\text{for } \beta_t^M = \left(\sum_i w_{it}^\pi (x_{it}, z_{it})' (x_{it}, z_{it}) \right)^{-1} \sum_i w_{it}^\pi (x_{it}, z_{it})' y_{it} \quad (8)$$

$$\text{i.e., } \hat{y}_t^M = \sum_i w_{it}^M y_{it} \text{ for}$$

$$w_{it}^M = w_{it}^\pi \left\{ 1 + \left((X_t, Z_t) - (\hat{x}_t^H, \hat{z}_t^H) \right) \left(\sum_i w_{it}^\pi (x_{it}, z_{it})' (x_{it}, z_{it}) \right)^{-1} (x_{it}, z_{it})' \right\} \quad (9)$$

The key to the method is the definition of the auxiliary variables. Let D be the set of units in the matched rotation groups (those with dwellings selected at both time points) at time t . Let y_{it}^* be the vector of key items for unit i at time t and Y_t^* the corresponding population totals. For $i \in D$, let $y_{t-1,i}^*$ be the previous month's value for the vector of key items, or if no value was reported let $y_{t-1,i}^*$ be imputed – I used $y_{t-1,i}^* = y_{t,i}^*$ as suggested by Singh (1996).

I look at modified regression estimates for z_{it} of the following form, for $a \in [0, 1]$:

$$z_{it} = (1-a) \frac{8}{7} y_{t-1,i}^* + a \left(y_{it}^* - \frac{8}{7} (y_{it}^* - y_{t-1,i}) \right) \text{ for } i \in D$$

$$= a y_{it}^* \text{ for } i \notin D. \quad (10)$$

Given this definition we have

$$\hat{z}_t^H = (1-a) \hat{y}_{t-1}^{+HD} + a (\hat{y}_t^{+H} - (\hat{y}_t^{+HD} - \hat{y}_{t-1}^{+HD})), \quad (11)$$

where $\hat{y}_{t-1}^{+HD} = 8/7 \sum_{i \in D} w_{t-1,i}^* y_{t-1,i}^*$ and $\hat{y}_t^{+HD} = 8/7 \sum_{i \in D} w_{t,i}^* y_{t,i}^*$ are estimates of Y_{t-1}^* and Y_t^* respectively based on units in D only and using this month's selection weights. For $a = 0$, \hat{z}_t^H is just the estimate \hat{y}_{t-1}^{+HD} . For $a = 1$, \hat{z}_t^H is this month's Horvitz-Thompson estimate minus an estimate of movement based on the matched rotation groups $\hat{y}_t^{+HD} - \hat{y}_{t-1}^{+HD}$. Values $a = 0$ and $a = 1$ give the methods MR1 and MR2 respectively of Singh *et al.* (1997). Use of an intermediate a was suggested by Fuller (1999).

An appropriate pseudo-benchmark Z_t would be an estimate of Y_{t-1}^* adjusted to agree with this month's weights. Following Singh *et al.* (1997) I used a step of generalised regression to adjust last month's modified regression estimator to add to this month's benchmarks:

$$Z_t = \hat{y}_{t-1}^{+M} + (X_t - \hat{x}_{t-1}^{+M}) \beta_t^{\text{adj}} \quad (12)$$

$$\text{for } \beta_t^{\text{adj}} = \left(\sum_i w_{t-1,i}^M x'_{t-1,i} x_{t-1,i} \right)^{-1} \sum_i w_{t-1,i}^M x'_{t-1,i} y_{t-1,i}^* \quad (13)$$

Note that $Z_t \approx \hat{y}_{t-1}^{+M}$ since $\hat{x}_{t-1}^{+M} = X_{t-1} \approx X_t$. This completes the definition of the modified regression estimators.

4.2 Properties of Modified Regression Estimators

The movement $\hat{y}_t^{+HD} - \hat{y}_{t-1}^{+HD}$ at (11) is actually based on the matched sample only (*i.e.*, units reporting at both times), since other units in the matched rotation groups D contribute zero to the movement (for the imputation used here). This may lead to the modified regression estimators having a lower sampling error than an AK estimator, as this "matched sample movement" is not affected by units not present in both months.

Unfortunately, this also gives the possibility of a bias if persons not represented in the matched sample have different behaviour to those in the matched sample. This may well be the case – the matched sample excludes persons that changed dwelling between the two months, and it is possible that changes of dwelling are related to changes of employment. This "matched sample bias" will be in addition to any time-in-survey bias.

Another problem arises with the MR2 estimator (*i.e.*, $a = 1$). If the k -th key variable $y_{it,k}^*$ has high month-to-month correlation then it will also have a high correlation with the k -th new auxiliary variable $z_{it,k}$. For such a

variable the element of β_t^M corresponding to $z_{it,k}$ will take some value y_i close to one. Using (7), (11), and $Z_t \approx \hat{y}_{t-1}^{+M}$, the MR2 estimator takes the form

$$\hat{y}_{t,k}^{+M} \approx (1 - y_i) \hat{y}_{t,k}^{+H} + y_i (\hat{y}_{t-1,k}^{+M} + (\hat{y}_{t,k}^{+HD} - \hat{y}_{t-1,k}^{+HD}))$$

$$+ \text{other terms.} \quad (14)$$

In this case it is possible that the matched sample movement at a given time will have a strong influence on estimates for many time points thereafter. In addition, any small bias in the movement will tend to accumulate over time. This danger was recognised by Fuller (1999), and referred to as "the drift problem". This was a motivation for his suggestion of the form of estimator given here, with a value of a less than 1.

In summary, modified regression has similar advantages to the AK composite weighting approach, but with possibly lower sampling error. The method is not difficult to apply, and avoids the need to separately calibrate the rotation groups to the benchmarks.

5. BEST LINEAR UNBIASED ESTIMATION (BLUE)

5.1 Fixed Window BLUE

The fixed window BLUE estimator (denoted by \hat{y}_t^B) is obtained by choosing an "optimal" linear combination of the rotation group estimates \hat{y}_t^{Rr} (as defined in 2.3) from a window of $l + 1$ months, as follows:

$$\hat{y}_t^B = \sum_{s=t-l}^t \sum_{r=1}^8 a_{sr} \hat{y}_s^{Rr} \quad (15)$$

where the parameters a_{sr} are chosen to minimise $\text{var}(\hat{y}_t^B)$ under the constraints $\sum_{r=1}^8 a_{sr} = 1$ for $s = t$ and $\sum_{r=1}^8 a_{sr} = 0$ for $s = t - l, \dots, t - 1$. These constraints ensure that \hat{y}_t^B will be unbiased for Y_t provided that the rotation group estimates are unbiased, *i.e.*, $E(\hat{y}_s^{Gr}) = Y_s$ for $s = t - l, \dots, t$.

The minimisation requires knowing the variances and covariances of the rotation group estimates. In practice these are estimated based on historical data. The problem can then be written in a matrix form: we aim to choose the column vector a (with elements a_{sr} for $s = t - l, \dots, t$ and $r = 1, \dots, 8$) so as to minimise a quadratic form $a'Va$ subject to constraints $C'a = c$. The relevant standard result (Rao 1973 page 65) is that the minimum occurs for $a = V^{-1}Cq$ where q is a solution of $(C'V^{-1}C)q = c$. In this study the matrix V was replaced by a correlation matrix, under the assumption that all the rotation group estimates in the window had the same variance.

5.2 Correlation Structure of Rotation Group Estimates

Since different correlation patterns give different BLUE estimates, choosing a correlation pattern has similar issues

associated with it as choosing parameters A and K in the AK composite. It is desirable to use the same linear combination for all estimates to assure additivity of the estimates.

I assumed a four parameter model for the correlation pattern:

$$\begin{aligned} \text{corr}(\hat{y}_t^{Gr}, \hat{y}_s^{Gr'}) &= \rho_{|t-s|}^W \quad \text{for } r-r'=t-s \\ &= \rho_{|t-s|}^B \quad \text{for } r-r'=t-s+8m \\ &\quad \text{for integer } m \neq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned} \quad (16)$$

Thus the correlation between estimates at lag k from the same rotation group is ρ_k^W if the rotation group contains the same dwellings at the two times, and ρ_k^B otherwise. Estimates from different rotation groups are uncorrelated. A four parameter model is used:

$$\rho_k^W = (1 - r_U^2)(\theta_P^k r_P^2 + \theta_B^k (1 - r_P^2)) \quad \text{and} \quad (17)$$

$$\rho_k^B = (1 - r_U^2) \theta_B^k (1 - r_P^2). \quad (18)$$

Bell and Carolan (1998) discusses this model. The parameter values used in this paper were $\theta_P = 0.87697$, $\theta_B = 0.94$, $r_U = 0.3101$ and $r_P = 0.90456$. These values result from fitting the model to estimated autocorrelations for rotation group estimates of proportion employed.

It is important to note that the BLUE estimates are unbiased regardless of the correctness of the assumed correlation model. The model used here aims to be optimal for estimates of employed persons, but turns out to perform well for unemployed persons as well. Trying other values for the model parameters did not give any marked improvement in standard errors for unemployed persons.

5.3 Improved BLUE Estimates

A problem with the BLUE estimates above is that GR estimates are required at rotation group level. The lower sample size at rotation group level may limit the benchmarks that can be used, as discussed for the AK. For the BLUE, however, an alternative approach is available.

The B1 estimator is defined by forming a BLUE estimator based on the Horvitz-Thompson estimators at rotation group level, and then applying the generalised regression technique to improve this estimator. This proceeds as follows. Define $y_{ti}^\# = a_{tR(t,i)} y_{ti}$ and $x_{ti}^\# = a_{tR(t,i)} x_{ti}$, where $a_{tR(t,i)}$ is the BLUE multiplier applicable to the rotation group unit i is in at time t . Then the BLUE estimator based on the Horvitz-Thompson estimators can be written

$$\hat{y}_t^{BH} = \sum_{s=t-l}^t \sum_i w_{si}^\pi y_{si}^\#. \quad (19)$$

Calibrating to the benchmarks we get the improved BLUE estimator B1:

$$\hat{y}_t^{B1} = \hat{y}_t^{BH} + (X_t - \hat{x}_t^{BH}) \hat{\beta} \quad (20)$$

$$\text{for } \hat{\beta} = \left(\sum_{s=t-l}^t \sum_i w_{si}^\pi x_{si}^\# x_{si}^\# \right)^{-1} \sum_{s=t-l}^t \sum_i w_{si}^\pi x_{si}^\# y_{si}^\#. \quad (21)$$

$$\text{i.e., } \hat{y}_t^{B1} = \sum_{s=t-l}^t \sum_i w_{si}^{B1} y_{si} \quad (22)$$

$$\text{for } w_{si}^{B1} = w_{si}^\pi a_{sR(s,i)} \{1 + (X_t - \hat{x}_t^{BH})\}$$

$$\left(\sum_{u=s-l}^s \sum_j w_{uj}^\pi a_{uR(u,j)} x_{uj}^\# x_{uj}^\# \right)^{-1} a_{sR(s,i)} x_{si}^\# \}^2. \quad (23)$$

Properties of the Blue and B1 Estimators

The BLUE and B1 estimates are sums of weighted unit data from a window of months. Each estimate needs only data from this window, and can be produced independently from the estimates for previous months – so the method is not recursive.

The same month of data will contribute with different weights to the estimate for different times. A unit will contribute a sizeable weight to its current month estimate, and a weight near zero, often negative, to estimates for other months. The work required in producing a table is the same as for GR multiplied by the size of the window. There is also a possibility of negative estimates for tiny cells containing no current units.

Note that in the B1 estimator the weights applied to months other than the current one are not forced to sum to zero. Under the model assumptions the estimate \hat{y}_t^{B1} remains unconditionally unbiased, since \hat{y}_t^{BH} and \hat{x}_t^{BH} are unbiased for Y_t and X_t respectively. In practice the current month contributes around 99.5 percent of the total weight. I consider the resulting bias to be small and not dangerous (leading as it does to some slight smoothing of the estimates over time).

For any estimate in which data from month to month is appreciably correlated, the BLUE and B1 estimates should have lower sampling error than the GR estimate. This is a theoretical advantage over a method that is designed for improving a predetermined set of estimates (like modified regression or the AK with composite weighting). In practice this advantage may not be too important, as for the LFS much of our interest is in a small number of well-defined estimates.

The user must also determine the time period or “window” from which estimates will be used. Using too many time points will be expensive computationally, while too few will limit the gains available. The seven month window used here was sufficient to obtain nearly all the available gains, while smaller windows give noticeably higher standard errors.

6. COMPARING THE METHODS

6.1 Method of Comparison

Estimates for July 1993 to January 1999 were produced based on data from January 1993 to January 1999. Estimates were obtained classified by month, state, sex, marital status and employment status. Estimates were also obtained for lag one movement, quarterly average and movement between successive quarterly averages.

Standard errors for these estimates were calculated using the "delete-a-group jackknife" technique (Kott 1998). The geographic units that form the first stage of sample selection were divided systematically into $G = 30$ groups, and "replicate groups" were formed consisting of the whole sample excluding the units from one of these groups. Each estimate studied was also produced for each of the G replicate groups. Writing e for the estimate and $e_{(g)}$ the estimate from replicate g , the delete-a-group jackknife estimate of standard error is given by

$$SE_{(e)} = \sqrt{\frac{G-1}{G} \sum_{g=1}^G (e_{(g)} - e)^2}. \quad (24)$$

Estimates and standard errors were obtained for each of the following estimators (listed with short mnemonics for later reference):

GR: Generalised regression estimate as currently used in the LFS

AK: AK estimate with $K=0.7$, $A = 0.06$

BL: BLUE based on 7 month window

B1: Improved BLUE based on 7 month window

MR2: MR2 estimator (modified regression with $a = 1$)

MF: Fuller's variant of modified regression ($a = 0.7$)

The modified regression estimators require a choice of the key variables to be optimised for. In producing the modified regression estimates in this report, z variables

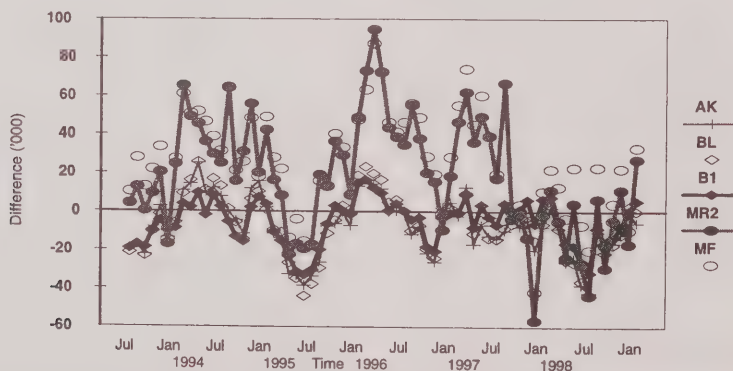
were produced for estimates of employed and unemployed for each state and sex. This gives a total of 32 extra auxiliary variables, in addition to the usual 560 post-stratum benchmarks used in generalised regression.

6.2 Differences From GR Estimate

The current GR estimator can be used as a basis of comparison for the other estimators. Rather than present graphs of level estimates, I present the differences of the alternative estimates from the current GR estimates. Graphs 1 and 2 show these differences for estimates of employed persons and unemployed persons respectively. To put the size of these differences in perspective, note that the published standard errors for the current estimate were 25,200 for employed persons and 7,900 for unemployed persons in January 1999 (and similar for other months).

The AK, BL and B1 estimates are quite similar, since in all three methods the contribution of a unit depends on its rotation group. In both graphs the AK, BL and B1 estimators appear to give lower values on average than the GR estimates. This indicates a change in the time-in-survey bias, resulting from putting less weight on the rotation group being sampled for the first time. The estimates vary up and down from their average difference for short periods.

The MR2 and MF estimates tend to be different to the other estimates since they emphasise the contribution of units from the matched sample. For employed persons, the MR2 and MF estimators are considerably larger on average than the GR estimates, up until September 1997. There is then a drop in the differences corresponding to the phase-in of a new sample from September 1997. For reasons that are not clear, over this period the matched sample behaved differently to the overall sample. This affects the difference between these modified regression series and the GR series. What may be of some concern is that the level change influences the level of the MR2 series for a considerable period thereafter, possibly a manifestation of the so-called "drift problem".



Graph 1. Difference of alternative estimates from GR estimate, employed persons ('000s), July 1993 to January 1999

For unemployed persons the M2 and MF estimates tend to be lower than the GR estimates. There is no evidence of a “drift problem” for unemployed persons, which is not surprising given the lower correlations involved.

6.3 Average Differences by Calendar Month

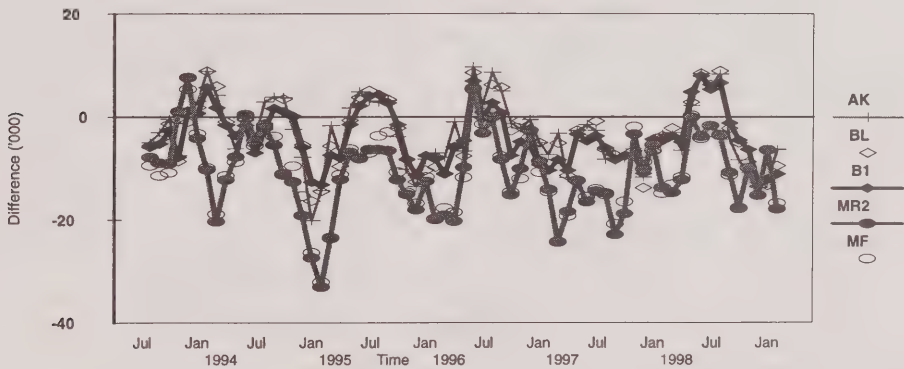
To quantify the likely change in bias from moving to a new estimator, the average difference over the period of each estimate from the GR estimate was calculated. It is possible that this difference is seasonal, so averages were obtained separately for each month of the calendar year, as well as overall. Average differences over the period July 1993 to January 1999 are given for employed persons in graph 3.

The graph shows that estimates of employed persons would have been higher on average using the MR2 or MF estimator. This upward difference for the modified regression estimators may actually be a feature of the particular period, since the difference has apparently dissipated since September 1997.

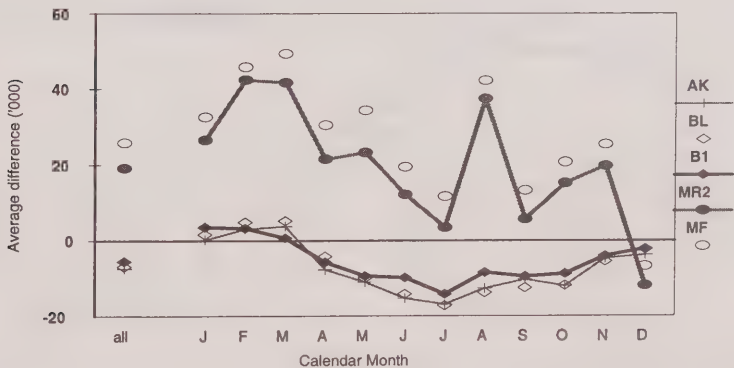
The other feature of the MR2 and MF estimates is that the difference for employed is highly seasonal. For example, the movement from December to January of the MR2 estimates is about 40,000 higher than the movement in the GR estimates. This suggests that the matched sample tends to miss people who were employed in December but not in January. The same seasonality shows up in looking at estimates from the matched sample directly. The matched rotation group movement does not show this large seasonal bias.

For the AK, BL and B1 estimates there is some seasonality in the differences, but the differences are much smaller.

Graph 4 shows the average differences of the various estimates from the GR estimate for unemployed persons over the same period. Here there appears to be a negative difference for all the estimators, though less pronounced for the AK, BL and B1 estimates than for the MR2 and MF. The change in seasonality from changing from the GR to the MR2 and MF estimators is again more extreme than for moving to the other estimators



Graph 2. Difference of alternative estimates from GR estimate, unemployed persons ('000s), July 1993 to January 1999



Graph 3. Average difference from GR estimate, overall and by calendar month, employed ('000)

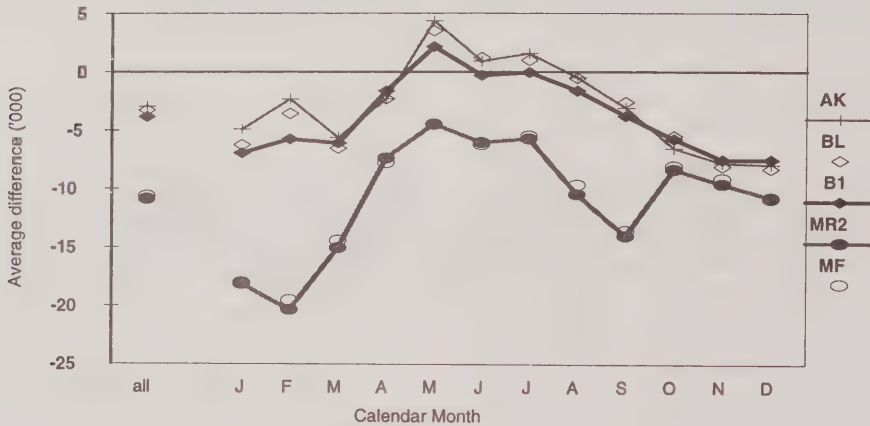
6.4 Standard Errors

Standard errors (SEs) of estimates overall, by marital status and by sex are presented in the following graphs. The SE estimates are obtained as a percentage of the SE estimate for the same estimate using the GR method (*i.e.*, the current LFS SEs), and these percentages are then averaged over the period for which they were produced (June 1993 to January 1999 for level estimates). Graphs 5, 6, 7 and 8 show SEs of both employed and unemployed persons for level, movement, quarterly average and movement of quarterly average respectively.

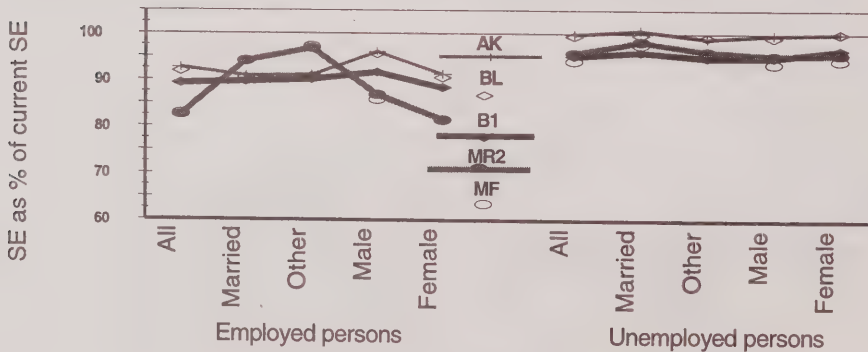
For all these estimates the BLUE-class estimator B1 has lower sampling error than the AK or BL estimators. Given that the B1 estimate appears to have similar bias and seasonality of bias it appears that the AK and BL estimators used in this study are not competitive with the B1 estimator.

The modified regression estimators MR2 and MF, on the other hand, give much lower sampling errors than the B1 estimator for employed persons for overall estimates and estimates by sex. These are key estimates used in the modified regression – other key estimates such as state estimates also gave similarly improved standard errors. Estimates by marital status are not key estimates, and these have higher standard errors for MR2 and MF than for the B1 estimator.

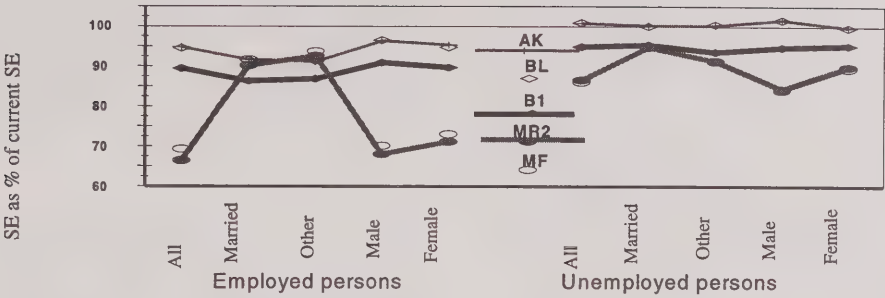
For unemployed persons the improvement in SEs from using MR2 and MF are less consistent, disappearing altogether for estimates of quarterly average. The B1 estimator is more consistent in lowering the standard error, although the gains available for unemployed are lower than for employed.



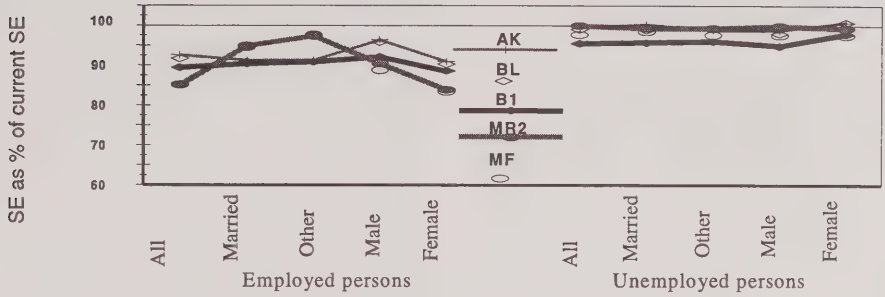
Graph 4. Average difference from GR estimate, overall and by calendar month, unemployed ('000)



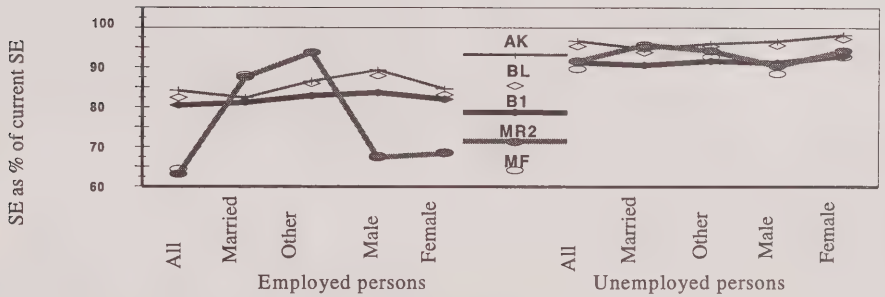
Graph 5. Standard Error of Level (% of current SE)



Graph 6. Standard error of movement (% of current SE)



Graph 7. Standard error of quarterly average (% of current SE)



Graph 8. Standard error of movement of quarterly average (% of current SE)

6.5 Seasonally Adjusted and Trend Series

The ABS uses the X11 package (Shiskin, Young and Musgrave 1967) to produce seasonally adjusted estimates that aim to remove various calendar effects from the series. The package also produces a trend, which is an indicator of the underlying behaviour of the series.

The trend value for a time point is revised as data for later times becomes available. I estimated the standard error of trend estimates at the end of the series (end trend) and for the same points when twelve further months of data are available (mid trend). Revisions of the trend (or trend movement) were defined as the difference between the mid and end values of the trend (or trend movement). The size

of the revision depends on the shape of the true series and on the sampling error in the estimated series. The mean squared trend revision for a series of unbiased estimates is the sum of two components: the mean squared trend revision that would have occurred even with no sampling error, and the variance of the estimate of revision. Thus the standard error of the revision is a measure of the sampling error component of the mean squared trend revision (see Bell 1999).

Seasonally adjusted figures are similarly subject to revisions. I present standard errors for level and movement of seasonally adjusted estimates at the end of the series. Standard errors for later revisions of these estimates were very similar.

The delete-a-group jackknife technique was used to produce estimates of standard error for the various trend and seasonally adjusted estimates. This technique requires producing replicate versions of the estimates. Unfortunately, the study provided replicate values for the time series only for time points from July 1993 to January 1999. Each of these replicate time series were supplemented by the previous 9 years of historical data so as to have sufficient data to apply the X11 package. Because the replicate seasonally adjusted and trend series are based on the same values before July 1993 the jackknife estimate of SE will tend to underestimate the true SE slightly, especially for times early in the series. To minimise this effect the measures of change in sampling error were averaged over months from January 1995 on only (and only up to January 1998, so that the 12 months to January 1999 can be used for estimating revisions).

Table 1
Standard error as percentage of standard error of
current GR estimator

	AK	BL	B1	MR2	MF
Employed persons:					
level	93	92	89	82	83
movement	95	95	89	66	69
quarterly average	93	92	89	85	85
movement of quarterly average	84	82	80	63	64
seasonally adjusted	94	92	90	87	88
movement of seasonally adjusted	96	95	91	68	71
trend at end	93	91	89	88	88
movement of trend at end	86	84	82	65	67
revision of trend	88	85	83	66	68
revision of movement of trend	89	86	84	67	69
Unemployed persons:					
level	100	99	95	96	94
movement	101	101	95	87	86
quarterly average	100	99	95	100	98
movement of quarterly average	97	95	91	92	90
seasonally adjusted	100	99	95	96	95
movement of seasonally adjusted	102	102	95	87	86
trend at end	99	98	95	99	97
movement of trend at end	97	95	92	93	91
revision of trend	97	95	91	91	89
revision of movement of trend	97	95	92	92	90

Table 1 gives these average standard errors for various seasonally adjusted and trend measures, relative to those available from the current GR estimator, for both employed and unemployed persons. Also in the table are corresponding figures for level, movement, quarterly average and movement of quarterly average, as presented in graphs 5 to 8.

I would argue that for many purposes the most important indicators are those that give the underlying direction of the series at the current end, *i.e.*, movement of quarterly average, and movement of trend. A reduced standard error for these items makes the underlying direction of the series at the end clearer, even for users who rely on visual inspection or on some smoothing process other than the

X11 trend. This in turn improves the ability to detect turning points in the underlying series.

For movement of trend the B1 estimator achieves an 18% reduction in standard error for employed persons and an 8% reduction for unemployed persons. For the MR2 these reductions are 35% and 7% respectively. The composite estimators also reduce the contribution of sampling error to revisions in the trend series.

6.6 Summary

This paper presents a variant of the BLUE estimator, the B1 estimator, which applies the generalised regression technique to a composite estimate based on a window of seven months of data. On Australian data, the B1 has lower sampling error than the traditional BLUE or AK estimators for a variety of measures including seasonally adjusted and trend estimates. The paper also evaluated a "modified regression" composite estimator MR2 proposed by A.C. Singh and a variant of this proposed by W. Fuller. These estimators gave considerably lower sampling errors than the B1 estimator for a number of measures, especially those based on employed persons.

The evaluation of a composite estimator will depend on many factors other than the sampling errors. The B1 estimator has the disadvantage that tabulations require weighted aggregation of seven months of data, whereas the modified regression estimators provide weights for a single month's data. On the other hand, the modified regression estimators may be biased if persons reporting in two successive months (the matched sample) are not representative of other persons (such as people moving house). Introducing the modified regression estimators would also induce a larger change in estimate and in seasonality than introducing the B1 estimator.

ACKNOWLEDGEMENTS

The author wishes to thank the referees for their very helpful input. This work was supported by the Australian Bureau of Statistics. Views expressed in this paper are those of the author and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted are used, they should be attributed clearly to the author.

REFERENCES

- BAILLAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of American Statistical Association*, 70, 23-29.
- BELL, P.A. (1998). Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates. Working Papers in *Econometrics and Applied Statistics*, Catalogue no. 1351.0, no. 98/2, ABS, Canberra.

- BELL, P.A. (1999). The impact of sample rotation patterns and composite estimation on survey outcomes. Working Papers in *Econometrics and Applied Statistics*, Catalogue no. 1351.0, no. 99/1, ABS, Canberra.
- BELL, P.A., and CAROLAN, A. (1998). Trend estimation for small areas from a continuing survey with controlled sample overlap. Working Papers in *Econometrics and Applied Statistics*, Catalogue no. 1351.0, no. 98/1, ABS, Canberra.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- FULLER, W.A. (1999). Canadian Regression Composite Estimation. Unpublished manuscript.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.
- JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural station research Bulletin*, 304.
- KOTT, P.S. (1998). Using the delete-a-group jackknife variance estimator in practice. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 483-486.
- LENT, J., MILLER, S. and CANTWELL, P. (1994). Composite weights for the Current Population Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 867-872.
- LENT, J., MILLER, S. and CANTWELL, P. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 130-139.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. Second edition, New York: John Wiley and Sons.
- SINGH, A.C., and MERKOURIS, P. (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.
- SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 120-129.
- SINGH, A.C., KENNEDY B., WU S. and BRISEBOIS F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.
- SHISKIN, J., YOUNG, A. and MUSGRAVE, J. (1967). *The X-11 variant of Census Method II Seasonal Adjustment*, Bureau of the Census, U.S. Department of Commerce, Technical Paper 15.
- YANSANEH, I.S., and FULLER, W.A. 1998. Optimal recursive estimation for repeated surveys. *Survey Methodology*, 24, 31-40.

Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation

JACK GAMBINO, BRIAN KENNEDY and MANGALA P. SINGH¹

ABSTRACT

The Canadian Labour Force Survey (LFS) is a monthly survey with a complex rotating panel design. After extensive studies, including the investigation of a number of alternative methods for exploiting the sample overlap to improve the quality of estimates, the LFS has chosen a composite estimation method which achieves this goal while satisfying practical constraints. In addition, for variables where there is a substantial gain in efficiency, the new time series tend to make more sense from a subject-matter perspective. This makes it easier to explain LFS estimates to users and the media. Because of the reduced variance under composite estimation, for some variables it is now possible to publish monthly estimates where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.

KEY WORDS: Rotating panel survey; Estimation system; Weighting; Change estimate; Level estimate.

1. INTRODUCTION

1.1 Why Composite Estimation?

The Canadian Labour Force Survey (LFS) is a monthly survey of 54,000 households selected using a stratified multistage design. Households stay in the sample for six consecutive months, thus five-sixths of the sample is common between consecutive months. Each month, the members of a selected household are asked questions about their labour force status, earnings, and so on. In the LFS estimation system used prior to 2000, initial design weights were modified using regression to produce final weights that respect age-sex and geographical (subprovincial region) *population control totals*. Each record then had a *unique final weight* that is used for all tabulations.

The estimation system used data from the current month only. No attempt was made to exploit the fact that the common sample can be used to improve estimates. However, characteristics such as employment by industry are highly correlated over time and unemployment is moderately correlated over time, thus there is potential for efficiency gains. Because of these gains, surveys similar to the LFS, such as the United States Current Population Survey (CPS), have used composite estimation to improve their estimates for many years. However, the LFS did not introduce composite estimation until January 2000.

In the early 1980s (see Kumar and Lee 1983), the CPS approach to composite estimation was studied for possible implementation in the LFS. Although the results showed that there were efficiency gains for Employed and, to a lesser extent, for Unemployed, it was felt that these gains were outweighed by the negative aspects of the method. These include the fact that the optimal parameters for Employed and Unemployed are quite different, which would have forced a trade-off between, on the one hand,

using a compromise set of parameters, thereby diluting the efficiency gains, and, on the other hand, having variables that do not *add up to totals* (e.g., Employed plus Unemployed would not equal Labour Force, unless one of the three is obtained as a residual). Another factor that worked against this form of composite estimation was that it was not *compatible with the existing weighting, estimation and dissemination systems* used by the LFS – the introduction of composite estimation would have required a complete overhaul of these systems.

Traditionally, the key estimates produced by the Labour Force Survey were monthly unemployment rates. However, with the increasing emphasis on estimates of employment level and on estimates of change in recent years, the need to find ways to make use of the common sample also increased since these estimates would benefit significantly. In the mid-1990s, therefore, interest in composite estimation was revived at Statistics Canada, and a regression-based method that fit in well with the existing LFS estimation system was developed. This method is described in Singh, Kennedy, Wu and Brisebois (1997) with a more up to date version included in Singh, Kennedy and Wu (2001). The new methodology allows for a choice of methods, depending on one's objectives. If the primary interest is in estimates of level, then one can use level-driven predictors in the procedure. If change is most important, then change-driven predictors can be used. One can go one step further and include both types of predictor in the procedure. However, in the latter case, the number of independent variables in the regression becomes large, which can lead to distortion of the final sample weights.

Preliminary results based on the new method using change-driven predictors and others using level-driven predictors were discussed at meetings of Statistics Canada's Advisory Committee on Statistical Methods. The method

¹ Jack Gambino, Brian Kennedy and Mangala P. Singh, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

addressed the problems with traditional composite estimators and showed substantial gains in efficiency. It was noted, however, that the estimator using change-driven predictors may lead to a drift in level estimates over time in some extreme situations. Also, it was decided, based on the committee's recommendation, that both estimates of level and of change should be given importance in the choice of predictors. After the exchange of technical notes between Wayne Fuller, J.N.K. Rao and Statistics Canada staff, a method suggested by Fuller, that combines the change-driven and level-driven approaches without the constraints associated with including both sets of predictors in the regression was adopted (see Fuller and Rao 2001). The solution is remarkably straightforward: take a linear combination of the level and change predictors: $X = (1 - \alpha)X_L + \alpha X_C$, and use it as the predictor. The change- and level-driven predictors are now special cases. Furthermore, one can choose α to reflect the relative importance one wishes to give to level versus change.

The present paper describes the new composite estimator in section 2. An extensive evaluation of this estimator was carried out using actual LFS data for a number of characteristics over a long period of time. The results of these studies are summarized in section 3. Unlike traditional composite estimators, the regression based composite estimator requires that the matching of the sample between two consecutive months be done at the individual record level. This creates some interesting situations where one has to deal with nonrespondents and in scope and out of scope individuals between two consecutive months in such away that the quality of estimates of change is not compromised. Section 4 discusses the imputation procedure developed to deal with various situations that arise when dealing with incomplete data for two consecutive months. Finally, the success of this new composite estimator is judged not only on its statistical efficiency but its stability over time and its cost effectiveness, while achieving the following objectives: (i) minimizing changes to the old estimation system, (ii) producing a unique weight for each sample unit (iii) respecting age-sex and geography control totals and (iv) producing consistent estimates (in the sense that, *e.g.*, Employed + Unemployed = Labour Force and Labour Force + Not In Labour Force = Population 15+). These objectives are discussed at various points in the paper, but especially in section 3.

2. THE REGRESSION COMPOSITE ESTIMATOR

Surveys such as the United States Current Population Survey have exploited their sample overlap by using K -composite or AK -composite estimators. Initially, the CPS used the K -composite estimator

$$y'_t = (1 - K)y_t + K(y'_{t-1} + \text{change}_{t-1,t})$$

with $K = 1/2$ for time t , where $\text{change}_{t-1,t}$ denotes an estimate of change based on the common, or matched, sample. This was later replaced by the AK -composite estimator

$$y'_t = (1 - K)y_t + K(y'_{t-1} + \text{change}_{t-1,t}) \\ + A(\text{unmatched} - \text{matched})$$

with $A = 0.2$ and $K = 0.4$ (see Cantwell and Ernst 1992). The optimal values of A and K depend on the variable of interest, and using different values for different variables poses problems of consistency (in the sense that parts do not add up to totals) in this approach. This prompted us to look for alternative approaches that satisfy the objectives mentioned at the end of the previous section.

It should be noted that we describe the new approach here at the person level, but in practice, person-level information is aggregated to the household level, and household-level records are then used by the estimation system.

To use regression for weighting in the old LFS estimation system, a regression matrix X is formed. Each person in the sample corresponds to a row of X . Each column of X corresponds to a control total; *e.g.*, column c may be Male 20-24, and the value in row i , column c will equal 1 if person i is a male between the ages of 20 and 24, and 0 otherwise (similarly for columns corresponding to geographical areas). For further details on the estimation methods used by the Labour Force Survey, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998).

To exploit the sample that is common between months, the X matrix is augmented by columns whose elements are defined in such a way that when this month's final weights are applied to the elements of each new column, the total is a composite estimate from the previous month, *i.e.*, last month's composite estimate is used as a control total (strictly speaking, the control total is based on weights that reflect the current month's population). As we noted in the introduction, there are several ways to define the new columns, depending on one's objectives. We present below only the alternatives that were proposed for implementation.

A typical new column will correspond to employment in some industry, say agriculture. If one is primarily interested in estimates of level, the following way of forming columns produces good results. Let M and U denote the matched (common) and unmatched (birth) sample, respectively. For person i , and times $t-1$ and t , let $y_{i,t-1}$ and $y_{i,t}$ be indicator variables which equal 1 whenever the person was employed in agriculture. Then let

$$x_i^{(L)} = \begin{cases} \bar{y}'_{t-1} & \text{if } i \in U \\ y_{i,t-1} & \text{if } i \in M, \end{cases}$$

where \bar{y}'_{t-1} is last month's composite estimate of the proportion of people employed in agriculture; in practice, we use $\bar{y}'_{t-1} = \hat{Y}'_{t-1}/P_{15+}$, where P_{15+} denotes the population aged 15 and over. The corresponding control total is last month's estimate of the number of people employed in

agriculture, *i.e.*, \hat{Y}'_{t-1} . Thus the end result is that the final weighted sum of the elements of the new column will equal last month's estimate. This is almost the same as forcing this month's weights, applied to last month's values for the common sample, to reproduce last month's estimate of employment in agriculture (after adjusting by 5/6). We have used the superscript L as a reminder that the goal here is to improve estimates of level.

If interest lies primarily in estimates of change, the following way of forming new columns of X produces good results:

$$x_i^{(C)} = \begin{cases} y_{i,t} & \text{if } i \in U \\ y_{i,t} + R(y_{i,t-1} - y_{i,t}) & \text{if } i \in M, \end{cases}$$

where R is a ratio that adjusts for the fact that five-sixths of the sample between months is common. The value $R = \sum_{\text{all}} w_i / \sum_{\text{matched}} w_i$ is used in the production system. For convenience, we used $R = 6/5$ during development since, in practice, the difference between the two is small because procedures to balance the weights by rotation group are used (*e.g.*, nonresponse adjustment is done separately by rotation group). As before, the corresponding control total is last month's estimate of the number of people employed in agriculture. Applying the final weights to the elements of this column of the X matrix and summing produces the equality

$$\hat{Y}'_{t-1} = \hat{Y}'_t - \hat{\Delta}_{t-1,t}^{M,f},$$

or, in words, last month's estimate equals this month's estimate minus an estimate $\hat{\Delta}$ of $Y_t - Y_{t-1}$ based on the common sample. We use the superscript f in $\hat{\Delta}$ as a reminder that the estimate of change is based on the final weights following composite estimation. In terms of the "pre-composite" weights, it is easy to show in the univariate case that

$$\hat{Y}'_t = (1 - b)\hat{Y}_t + b(\hat{Y}'_{t-1} + \hat{\Delta}_{t-1,t}^M),$$

where b is the regression coefficient and $\hat{\Delta}$ is the estimate of change based on the original weights. The more general case where auxiliary variables are present is given by Fuller and Rao (2001, equation 2.3).

Earlier results have shown that using the L controls produces better estimates of level for the variables added to the X matrix as controls. Similarly, adding C controls produces good estimates of change for the variables that are added. Singh, *et al.* (1997, 2001) present efficiency gains for C -based estimates of level and change and refer to earlier results on L -based estimates.

Early in the development, an estimation system that used only the C -based controls was considered. However, there was some concern expressed about an estimation system based solely on change-driven controls since estimates of level are also very important (for example, they play a key role in the federal government's Employment Insurance program). These concerns are summarized in Fuller and Rao (2001).

In principle, we can add both L and C controls to the regression, but this would result in a large number of columns in the X matrix, which has undesirable consequences such as an increased number of extreme final weights, including negative weights. To avoid this, we would have to limit the number of industries included in the estimator. Wayne Fuller (see Fuller and Rao 2001) proposed an alternative which allows us to include the industries of greatest interest while allowing us to compromise between improving estimates of level and improving estimates of change. Fuller's alternative is to take a linear combination of the L column and the C column for an industry and use it as the new column in the X matrix, *i.e.*, use

$$x_i = (1 - \alpha)x_i^{(L)} + \alpha x_i^{(C)}.$$

The original level- and change-driven variables are special cases of Fuller's compromise.

Choice of α : Fuller and Rao (2001) showed that, based on some reasonable assumptions, values of α such as 0.65 and 0.75 produce reasonable estimates of both level and change. The actual choice of α depends on the variable of interest (specifically, its correlation over time) and on the relative importance of level versus change.

Our studies (see Appendix 1) showed that for the two most important variables, employed and unemployed, the best choices of α for estimates of level are 0.39 and 0.24, respectively. The corresponding values for estimates of change are 0.99 and 0.81, respectively. Clearly, there is a need to compromise between the goals of improving estimates of level and estimates of change.

To decide which values of α to study, we obtained compromise values of α by averaging the level-driven and change-driven values for each variable, *i.e.*, we obtained approximately 0.7 and 0.52 for employed and unemployed, respectively. Results based on the values $\alpha = 1$ and $\alpha = 0.75$ had already been produced, so we added results for $\alpha = 0.67$ and $\alpha = 0.6$. Based on the results discussed below, which show that there are no substantial differences in the results for the three values 0.6, 0.67, and 0.75, we chose to implement the value $\alpha = 2/3$ in the production system.

3. FEATURES, PROPERTIES AND RESULTS

We present a summary of some of the features and properties of the regression composite estimator. Some graphical and numerical results are presented in section 3.1 below.

Systems implementation. An important advantage of the estimator is that it can be implemented within the old LFS estimation system in a straightforward manner since, essentially, one needs to augment the regression matrix, as described above. This was an important factor in our initiative to study and finally introduce composite estimation as

otherwise it would have cost a great deal more to change the system.

Weighting. Unlike the *A-K* estimator, where weighting to satisfy population control totals and composite estimation are separate steps, weighting for the regression composite estimator is done in one step, *i.e.*, simultaneously with weighting to satisfy the age-sex and geographical controls. For illustration, the way the regression matrix would be augmented when elements $x_i^{(C)}$ defined in section 2 are added is shown in Appendix 3. Adding the elements $x_i = (1 - \alpha)x_i^{(L)} + \alpha x_i^{(C)}$ is similar. This not only preserves the consistency mentioned next but also retains the benefits of the controls applied to the usual regression estimator, *i.e.*, the age-sex and geographic controls in our case.

Consistency. Because weighting for age-sex and geographical controls is done at the same time as weighting for the composite estimate controls, consistencies are preserved. In particular, parts add up to totals; *e.g.*, Employed + Unemployed = Labour Force. In other approaches to composite estimation, consistency is achieved by other means which require either a separate step or a compromise of some kind.

Efficiency gains. For the variables that are added as control totals, there are substantial gains in efficiency for both estimates of level and of change. For $\alpha = 1$, the gains for estimates of change can be dramatic; by choosing a smaller value of α we gain more for estimates of level while reducing the magnitude of the gains for estimates of change. Some results for the case $\alpha = 2/3$ are given in section 3.1.

Seasonal adjustment. The time series of employment by various industries are scrutinized by both internal and external users of the Labour Force Survey. One important consequence of the gain in efficiency is that several of these series which could not be seasonally adjusted in the past can now be seasonally adjusted. In other words, composite estimation increases the signal-to-noise ratio sufficiently that seasonal adjustment becomes effective. A related consequence of composite estimation that is popular with users is that several estimates that were published as three-month moving averages are now published as monthly estimates.

Systematic differences between composite and usual level estimates. In theory, the expectations, taken over all possible samples, for both the usual and composite estimators should be the same, making them both unbiased or almost unbiased. One would therefore expect that the estimates of level obtained using the two estimators would criss-cross each other over time. In practice, however, this does not happen. This is due to the fact that, when actual survey conditions are taken into account, the composite estimator and the usual estimator do not have the same expected value; for example, see Bailer (1975) and Kumar and Lee (1983) for results on the *K*- and *AK*-composite estimator, respectively. Kumar and Lee show this by deriving explicit expressions for the expected value of the

usual estimator and the *AK*-composite estimator. The matched and unmatched samples differ because of differences in nonresponse rates and the mode of data collection (*e.g.*, personal versus telephone interviewing, centralized versus decentralized interviewing). In practice, the units in the "birth" sample have a higher nonresponse rate, and the missing households tend to be smaller and have higher employment rates than the responding ones. Since the usual estimator and the composite estimator give different weights to the matched and unmatched sample, they will have different expected values. Thus time series for the two estimators can display systematic differences. In practice, these differences are usually swamped by sampling variation, but they become evident for more precise series such as Employed for big provinces like Ontario and for Canada. Our results for Employed are consistent with those described by Bailer (1975) for the U.S. Current Population Survey, *i.e.*, the composite estimates for Employed tend to be smaller than the usual estimates. For Unemployed in Ontario, the difference between the two types of estimates tends to be much smaller.

Ways of reducing systematic differences between estimates from different rotation groups are currently being investigated. In particular, the possibility of introducing a weight adjustment for the number of households of different sizes by rotation group is being studied as a way of adjusting for the fact that small households are under-represented in the birth rotation. This would benefit both the composite estimators and the usual regression estimator, and would probably reduce the gap between them.

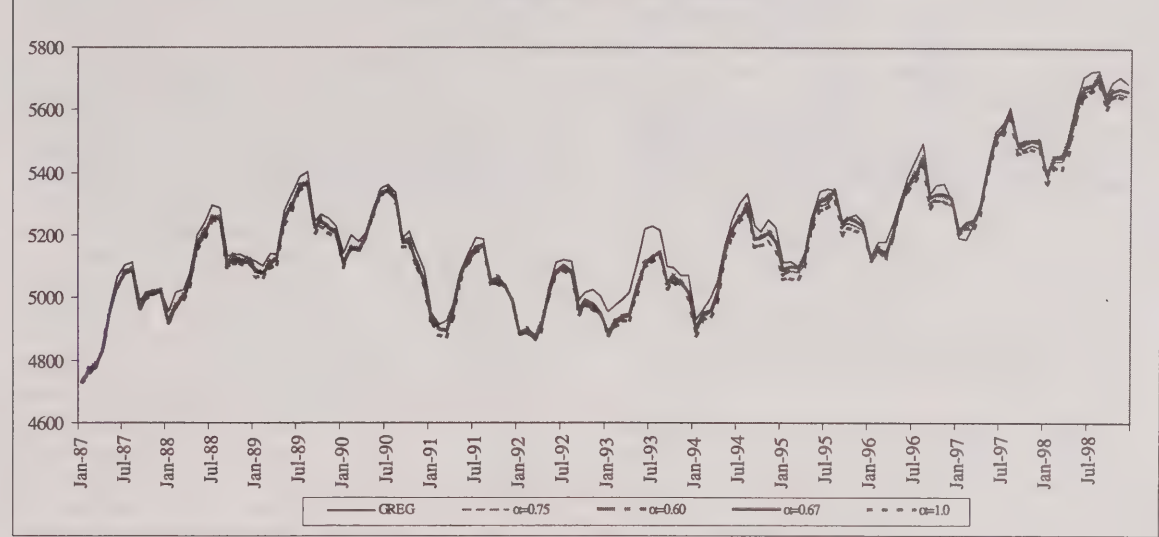
3.1 Empirical Results

Employment and unemployment at the provincial level. Graph 1 shows total employment at the province level from 1987 to 1998 for Ontario. The time series for the composite estimation series for the four values of α , *i.e.*, for 0.6, 0.67, 0.75 and 1 behave similarly. In these graphs, it is clear that there is a change in level for this series under composite estimation – the estimated number of employed persons is lower. The seasonally adjusted versions of the Ontario employment series based on the usual estimator and on the composite estimator for $\alpha = 0.67$ are shown in Graph 2.

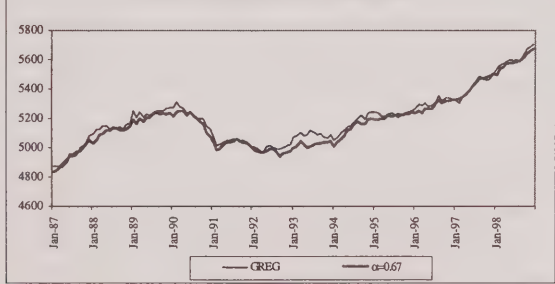
Graph 3 compares the usual estimates of Ontario unemployed to the regression composite estimate for $\alpha = 0.67$. The effect of composite estimation on this variable is clearly less pronounced than on employment-related variables.

Graph 4 compares year-to-year changes in Ontario employment for the two estimators. Each point in the series is the difference between employment in year y , month m and year $y - 1$, month m . For example, the first point is January 1988 employment minus January 1987 employment. The composite estimation series is clearly smoother, especially in the second half of the twelve-year period.

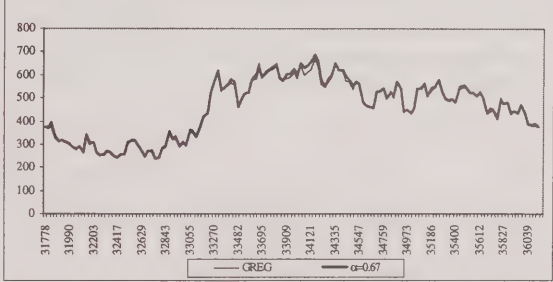
Graph 1: Ontario Employed (000's) Unadjusted



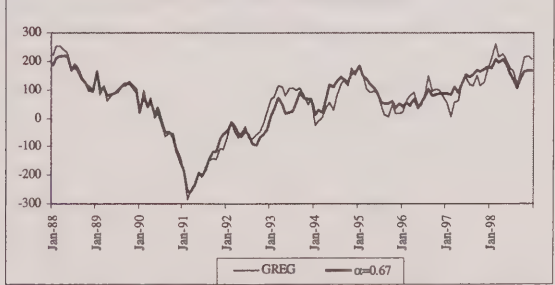
Graph 2: Ontario Employed (000's) Seasonally Adjusted



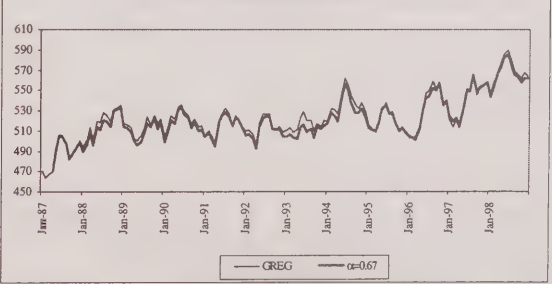
Graph 3: Ontario Unemployed (000's) Unadjusted



Graph 4: Ontario Year-to-Year Change in Employment



Graph 5: Employed in Economic Region 510



Employment by subprovincial region. Graph 5 compares the usual estimate of employment with the composite estimate with $\alpha = 0.67$ for an economic region in Ontario. The results for other subprovincial regions are similar. The behaviour of the usual and composite estimate series are very similar, thus, the effect of composite estimation is

neither beneficial nor harmful. For special tabulations, the LFS estimation system has the flexibility to allow the user to add controls at the economic region level if needed. There is already a control for the total population in each economic region.

Employment by industry, and seasonal adjustment.

The composite estimates were compared to the usual regression estimate for sixteen industries. Graph 6A-6D show the results for two of them in Ontario. Though not included in these graphs, once again, the four values of α result in composite estimation series that generally behave similarly, although sometimes the series for $\alpha = 1$ departs from the others. The composite estimation series tend to be less volatile than the regression series. This is particularly noticeable for the seasonally adjusted Trade series which we have included here because it illustrates the most extreme case. For this series, the behaviour of the original regression estimates in the first few years, in both the seasonally adjusted and unadjusted series, is difficult to explain from a subject-matter viewpoint. The behaviour of the Manufacturing series is more typical of the remaining fourteen industries.

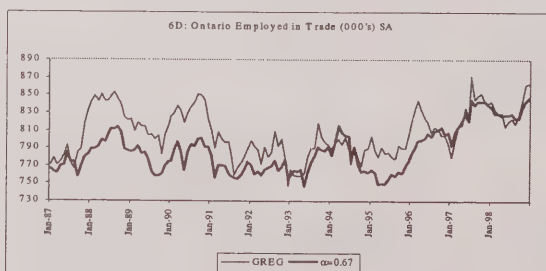
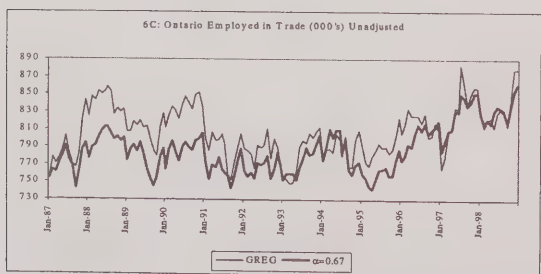
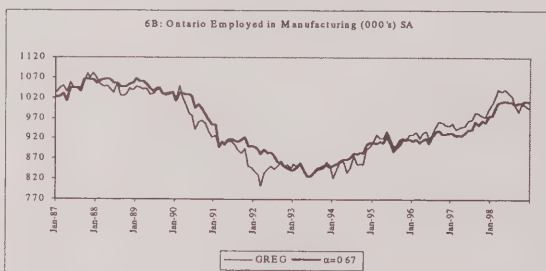
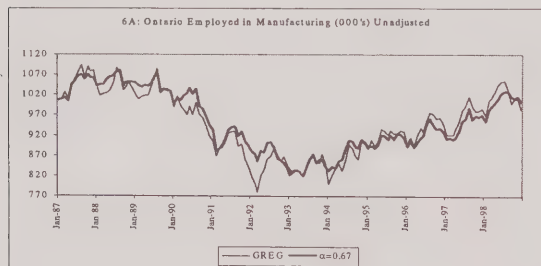
Comparing the seasonally adjusted (Graph 6D) and unadjusted (Graph 6C) series for Trade, we see that seasonal adjustment has had relatively little effect on the regression series, but has changed the composite series significantly. This is a manifestation of the ability of composite estimation to increase the signal-to-noise ratio sufficiently to make seasonal adjustment effective.

The seasonal adjustment program used by the Labour Force Survey computes a variety of measures that are used as indicators of the effectiveness of seasonal adjustment. Some of these measures are presented in Appendix 2. These show that, for Ontario employment in the twelve-year

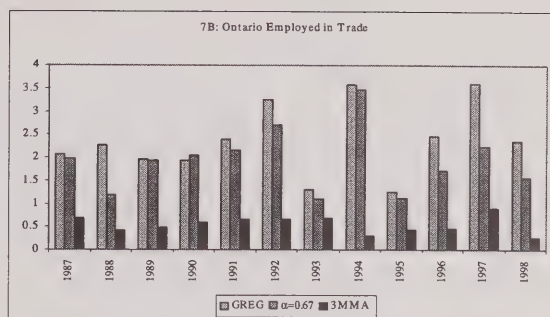
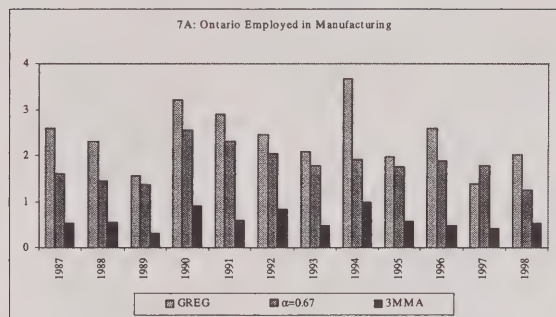
period 1987-1998, composite estimation increases the number of industries that can be successfully seasonally adjusted. Results for other provinces and for Canada as a whole are similar.

A measure of stability. For several important data series, instead of monthly estimates, three-month moving averages were published in the past. This was due to the high sampling variability associated with these series, leading to unacceptable volatility in the monthly series. Of particular interest are province-level estimates by industry and by class of worker. It had been anticipated that the composite estimates for these series would demonstrate more stability, allowing the publication of monthly estimates instead of three-month averages. A measure of stability, the index of volatility, is computed as follows. For each industry, the month-to-month change in employment is calculated from seasonally adjusted estimates. Then the difference between consecutive change estimates is computed. The absolute value of this "change in the change" is expressed as a percentage of the corresponding monthly total estimate. These percentages are then averaged over the entire year. Large values of this measure occur when a series has many consecutive movements in opposite directions, indicating volatility.

The volatility index was computed for sixteen industries. Graphs 7A and 7B for two of these industries, Ontario Manufacturing and Trade, are included here, comparing the usual estimator, the three-month moving average of the usual estimator and the monthly composite estimator with



Graph 6. Selected Employment by Industry



Graph 7. Index of Volatility

$\alpha = 0.67$. For Manufacturing, the average indexes for the usual, composite and moving average estimates are 2.4, 1.8 and 0.60, respectively. For Trade, the corresponding values are 2.4, 1.9 and 0.55. For all industries, the volatility of the composite estimates typically falls between that of the usual monthly and three-month average estimates. Occasionally, for isolated years, the composite estimates are less volatile than the three-month averages or more volatile than the usual monthly estimates, but generally the volatility of the composite estimates is between that of the usual monthly estimates and that of the three-month moving averages. We also note that when the usual monthly estimates exhibit extreme volatility, the composite series tend to be more stable. The monthly regression estimates compete with the composite estimates only when the volatility index is low for both of them.

With the introduction of composite estimation, three-month moving averages were dropped in favour of the more desirable monthly estimates for industry series.

Variance estimates. For variables that are added as control totals, such as employment by industry, there can be substantial gains in efficiency at the province level, where efficiency is defined as $\text{Var}(\text{greg})/\text{Var}(\text{composite})$. For most industries, gains of 10 to 20 percent are typical, but they can be as high as 40 percent. A 40 percent efficiency gain corresponds, for example, to reducing a 15 percent coefficient of variation to 12.7 percent and a 10 percent coefficient of variation to 8.5 percent. For province-level employment and unemployment estimates, the efficiency gains are more modest, typically in the five to ten percent range. For estimates of month-to-month change, the efficiency gains for controlled variables are bigger, usually more than double the gains for estimates of level.

For variables that are not controlled, there is little or no effect of composite estimation on efficiency unless the variable is highly correlated with a controlled variable. For example, at the province level, Employed Males shows a gain in efficiency because it is correlated with total employed, which is controlled. On the other hand, employment by subprovincial economic region shows neither gains nor losses.

4. TREATMENT OF MISSING DATA

By definition, the x_i variables involve data from the current and previous month. This leads to complications when, for a given person in the common sample, data is available only for one month. This may occur due to non-response in either month or when a move or change in scope has taken place between the two months. The different cases that may occur are represented in the following diagram, where R denotes a response, X denotes a nonresponse and O denotes a unit that is out of scope.

	A	B	-	C	D
Month t	XXX...	RRR...	RRR...	RRR...	OOO...
Month $t-1$	RRR...	XXX...	RRR...	OOO...	RRR...

In all these cases, namely A, B, C, and D, the objective is to find a solution such that $\sum_{i \in S} w_{it} x_{it}$ is still an estimator of Y_{t-1} . We set the following two objectives for handling the situation of missing data from either month of the common sample:

- retain as many valid responses as possible, *i.e.*, the option of removing a unit from the estimation process is rejected
- develop an imputation method that does not understate the estimate of change in any significant way.

In the case of nonresponse, there are two situations: Case A, where a household responded last month but not this month, and Case B which is the reverse situation. In the following, i denotes a person in an affected household.

Case A: Replace y_{it} by \hat{y}_{it} . This can be achieved in a number of ways. A simple approach is to replace y_{it} by the corresponding response from the previous month, *i.e.*, $y_{i, t-1}$. During the early stages of the study, this approach was used but rejected later as it can bias (understate) the estimate of change significantly. For the LFS estimation system, it was decided to use the previous month's known demographic and employment characteristics of persons to

form imputation classes and then use hot deck imputation (*i.e.*, current month's data) to obtain \hat{y}_{it} . An alternative would be to use a mean of some sort.

Case B: The procedure is analogous, *i.e.*, when last month's value is missing, then imputation classes are formed using data from month t and the donor is found using data from responding units in month $t-1$.

In the case where unit i has moved or changed scope, the following situations may arise.

Case C: Suppose that unit i was out of scope at time $t-1$ but is in scope at time t (*e.g.*, a person who just turned 15, or a newly arrived immigrant). Then unit i should contribute 0 at time $t-1$ and y_{it} at time t . Hence we let $x_{it} = 0$ since $\sum w_{it} x_{it}$ should estimate Y_{t-1} .

Case D: Conversely, suppose that unit i was in scope and is now out of scope. This includes, *e.g.*, people who left the country, joined the military or died. Such units should be dropped since the target population is the in-scope population at time t (and the ultimate goal is to estimate Y_t). Since we sample dwellings but collect data for individuals within those dwellings, two other situations arise due to movement of persons in and out of the sampled dwellings.

Case i): Suppose that unit i was in the population at both times but in a sampled dwelling only at time t (*i.e.*, a person who moved from a non-sampled dwelling to a sampled dwelling). Then his/her status at time $t-1$ is unknown, *i.e.*, $y_{i,t-1}$ is unknown. For all such cases, as in the nonresponse case, we can impute a value $\hat{y}_{i,t-1}$ for $y_{i,t-1}$ either from a donor in the sample or by a sample mean. The LFS uses hot deck imputation.

Case ii): Finally, consider the case where unit i was in the sample at time $t-1$ but moved to a non-sampled dwelling at time t . Since the LFS sample is a sample of dwellings and not a sample of people, this unit should simply be dropped when computing estimates of Y_t .

5. CONCLUSION

The composite estimator described in this document meets all the objectives that were set at the beginning of this project and summarized in the introduction. It produces estimates of level and change that are more efficient than the estimates produced by the usual regression estimator while satisfying all operational and consistency constraints. The impact of the composite estimator with the value $\alpha = 2/3$ on the many time series produced by the Labour Force Survey is generally moderate. When the impact is substantial, as in the Ontario Trade series, for example, the new series tend to make more sense from a subject-matter expert's perspective. This type of improvement in the series makes it easier to explain LFS estimates to users and the media.

The composite estimates have other features that users find very desirable. Because of the reduced variance under composite estimation, it is possible to publish monthly

estimates in many cases where only three-month moving averages were published in the past. In addition, a greater number of series can be successfully seasonally adjusted.

Implementation of composite estimation for the LFS is an important first step. Studies to improve the treatment of nonsampling errors are ongoing, and their results can be incorporated into the weighting and estimation system at any time. The system has the great advantage that it is very flexible. For example, the value of α can be changed easily, hence a comparison of a broad range of α values for a number of important variables is planned. This may lead to a system in which different α values are used for different control variables, while still having a unique final weight per record.

ACKNOWLEDGEMENTS

We would like to thank Avi Singh and Statistics Canada's Advisory Committee on Statistical Methods for their contributions to this project. We are also grateful to the many people whose comments on earlier versions of this paper improved it greatly.

APPENDIX 1

Relationship between α , ρ and (A, K) . Kumar and Lee (1983) found optimal values of A and K in AK -composite estimation for estimates of level and change as a function of the correlation coefficient ρ . We derived an approximate relationship between the A and K values, ρ and α . This result was then used to find good values of α for several variables. These are presented in Tables 1 and 2 for estimates of level and change, respectively. The A and K values in the tables are the optimal ones for the corresponding value of ρ . The values of α in the tables are consistent with those obtained by Wayne Fuller based on an AR(1) model (personal communication). The value of α for Labour Force in Table 2 exceeds one because of the approximation.

Table 1
 α Values for Several Variables – Level

Variable	ρ	A	K	α
Employed	0.852	0.49	0.8	0.385
Unemployed	0.58	0.38	0.5	0.242
Labour Force	0.843	0.48	0.8	0.403
E.P. Agriculture	0.955	0.38	0.8	0.448

Table 2
 α Values for Several Variables – Change

Variable	ρ	A	K	α
Employed	0.852	0.1	0.9	0.995
Unemployed	0.58	0.2	0.6	0.806
Labour Force	0.843	0.1	0.9	1.009
E.P. Agriculture	0.955	0	0.9	0.959

APPENDIX 2:
Seasonal adjustment measures for Ontario employment by industry

Industry	F Value			M7			SMOOTH	
	greg	$\alpha = 0.60$	$\alpha = 0.75$	greg	$\alpha = 0.60$	$\alpha = 0.75$	greg	$\alpha = 0.60$
Agriculture	87.76	120.18	112.7	0.27	0.23	0.24	37.94	45.36
Forestry	21.34	24.58	23.22	0.5	0.52	0.57	21.76	26.78
Utilities	4.29	3.48	6.8	1.1	1.25	0.82	15.39	15.52
Construction	128.3	275.06	246.93	0.26	0.16	0.17	41.68	57.5
Manufacturing	38.22	55.6	69.21	0.37	0.3	0.3	29.02	31.94
Trade	9.93	15.12	20.35	0.8	0.68	0.53	25.13	34.92
Transportation	9.16	8.64	9.69	0.94	0.75	0.7	15.36	23.33
Finance	6.49	8.94	8.84	1.22	0.76	0.77	13.45	19.67
Professional	5.3	12.91	9.81	1.03	0.72	0.76	12.45	19.52
Management	14.72	24.98	20.35	0.67	0.52	0.52	16.2	22.17
Education	67.37	219.62	214.37	0.33	0.16	0.19	53.25	66.47
Health Care	8.78	10.73	8.48	0.8	0.66	0.75	16.09	19.92
Information	21.13	52.31	62.94	0.66	0.36	0.35	24.29	33.46
Accommodations	44.85	75.37	78.03	0.36	0.34	0.3	31.89	44.29
Other Services	2.61	13.17	12	1.41	0.75	0.81	18.58	26.27

Description of Measures

F-value: F-value for the test performed within the X11-ARIMA program to check for the presence of stable seasonality. The higher the F, the more significant is the presence of stable seasonality.

M7: Measure that combines the test for stable and moving seasonality. Generally, when M7 is greater than 1, there is no identifiable seasonality present in the series; therefore, the series should not be adjusted.

SMOOTH: Percentage difference between the standard deviation of the month-to-month changes in the original series and the standard deviation of the month-to-month changes in the seasonally adjusted series. The larger this value, the more smoothing was obtained through the seasonal adjustment process.

APPENDIX 3: Implementing Regression Composite Estimation within the LFS Estimation Framework: Illustrated Using the Change-driven Approach

Original X matrix

Age-sex indicators							Region indicators						
0	0	1	0	.	.	.	0	0	1	0	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	0
.
.
X_1	X_2	X_k	X_{k+1}	X_p

Population control totals

←

Modified X matrix for composite estimation when $x_i^{(c)}$ are added

Age-sex indicators								Region indicators								E	U	Ag	mining	services		
0	0	1	0	.	.	.	0	0	1	0	.	.	.	0	a	0	0	b	.	.	.	0
0	1	0	0	.	.	.	0	0	1	0	.	.	.	0	c	0	d	0	.	.	.	0
.
.
X ₁	X ₂	X _k	X _{k+1}	X _p	E'	U'	Ag'	S'

E' is last month's
employment estimate

For *birth* units, set a, b, c, \dots to indicate this month's status (e.g., $a=1$ if employed, 0 otherwise). For *matched* units, do the following:

$a = e_t + (e_{t-1} - e_t) \times 6/5$ where $e=1$ if person is employed, $e=0$ otherwise

$d = ag_t + (ag_{t-1} - ag_t) \times 6/5$ where $ag=1$ if person is employed in agriculture, $ag=0$ otherwise

Examples:

- (i) Suppose Person 2 was employed in agriculture both last month and this month. Then $e_{t-1} = e_t = 1$ and $ag_{t-1} = ag_t = 1$, hence $c = 1 - 0 = 1$ and $d = 1 - 0 = 1$.
- (ii) Suppose Person 2 was employed in agriculture last month and in mining this month. Then $e_{t-1} = e_t = 1$, $ag_{t-1} = 1$ and $ag_t = 0$ hence $c = 1 - 0 = 1$ and $d = 0 + (1 - 0) \times 6/5 = 1.2$.
- (iii) Suppose Person 2 was employed in mining last month and in agriculture this month. Then $e_{t-1} = e_t = 1$, $ag_{t-1} = 0$ and $ag_t = 1$ hence $c = 1 - 0 = 1$ and $d = 1 + (0 - 1) \times 6/5 = -0.2$.

REFERENCES

- BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- CANTWELL, P.J., and ERNST, L.R. (1992). New developments in composite estimation for the Current Population Survey. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, 121-130.
- FULLER, W.A., and RAO, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.
- GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. and LINDEYER, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue number 71-526.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 178-201.
- SINGH, A.C., KENNEDY, B. and WU, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, 27, 33-44.
- SINGH, A.C., KENNEDY, B., WU, S. and BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

Variance Estimation After Imputation

JAE-KWANG KIM¹

ABSTRACT

Imputation is commonly used to compensate for item nonresponse. Variance estimation after imputation has generated considerable discussion and several variance estimators have been proposed. We propose a variance estimator based on a pseudo data set used only for variance estimation. Standard complete data variance estimators applied to the pseudo data set lead to consistent estimators for linear estimators under various imputation methods, including without-replacement hot deck imputation and with-replacement hot deck imputation. The asymptotic equivalence of the proposed method and the adjusted jackknife method of Rao and Sitter (1995) is illustrated. The proposed method is directly applicable to variance estimation for two-phase sampling.

KEY WORDS: Two-phase sampling; Item nonresponse; Deterministic imputation; Random imputation.

1. INTRODUCTION

Imputation, inserting values for missing items, is commonly used for handling missing survey data. An advantage of imputation is its convenience. That is, we can apply standard complete data programs for computing point estimates to the imputed data set. Rubin (1996), Fay (1996), and Rao (1996) reviewed various issues on imputation.

All imputation methods use some type of model. After designating a model, we can use either deterministic imputation or random imputation based on the model. Under random imputation, missing values are imputed by the use of some form of probability sampling. We call this additional random mechanism the imputation mechanism. On the other hand, deterministic imputation does not introduce an additional random mechanism. When the set of respondents is viewed as a random sample from the original sample, the selection mechanism of the respondents is called the response mechanism. The response mechanism is often regarded as the second phase of sampling. See Särndal and Swensson (1987) for details.

With a suitable imputation model and method, the bias due to nonresponse can be greatly reduced relative to using only the observed data. However, it is well known that a variance estimator which uses the imputed data as if it were observed data is inconsistent.

Various methods have been proposed for variance estimation after imputation. Rubin and Schenker (1986) and Rubin (1987) advocate multiple imputation. Multiple imputation creates multiple data sets and calculates the complete data statistics for each imputed data set. The variance estimator is calculated by combining two terms, the within-dataset variance term and the between-dataset variance term. Multiple imputation applies standard variance estimators to each data set to compute within-dataset variance terms and applies the standard point estimators to compute

a between-imputed-dataset variance term. This method requires the imputation method to be proper. That is, the imputation should satisfy conditions 1-3 in Rubin (1987, pages 118-119). These conditions are not always easy to achieve. (For example, see Fay 1992). Even the multiple imputation methods described in Schafer (1997) are not shown to be proper in the sense of Rubin. As noted by Rao (1996), some commonly used imputation methods, including hot deck imputation and regression imputation, are not proper.

Rao and Shao (1992) and Rao and Sitter (1995) proposed an adjusted jackknife variance estimator. The suggested procedure is applicable to a number of imputation methods and sample designs. The actual calculation using standard complete data software is not easy because special computations are performed to adjust the imputed values for each pseudo replicate. Also, Särndal (1992) proposed a variance estimation method that explicitly uses the model considered for imputation.

Essentially, Rubin's method generates several pseudo data sets for variance estimation and applies the standard variance estimators to each data set to compute the within-dataset variance terms, while Rao's method and Särndal's method apply a special variance estimator to the imputed data set. In this paper, a method to create a single pseudo data set for variance estimation is proposed. In section 2, the new method is introduced in a two-phase sampling set-up. In section 3, we illustrate extensions of the suggested method to the random imputation method. In section 4, we extend the suggested method to complex sampling designs. In section 5, comparisons are made with the adjusted jackknife variance estimator. In section 6, a limited simulation study is presented. Some concluding remarks are made in section 7. Outlines of some proofs are given in the appendix.

¹ Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

2. A VARIANCE ESTIMATION METHOD

We outline a variance estimation procedure applicable for two-phase samples and for imputed samples. The procedure requires a separate data set for variance estimation in addition to the tabulation data set. To introduce the procedure and to illustrate the concepts, consider a two-phase sample. Let the second phase be a simple random sample of size r selected from the first phase, which is a simple random sample of size n selected from an infinite population. Let the regression estimator of the mean of a characteristic y be

$$\hat{\mu}_y = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}, \quad (1)$$

where

$$(\bar{y}_2, \bar{x}_2) = r^{-1} \sum_{i=1}^r (y_i, x_i),$$

$$\bar{x}_1 = n^{-1} \sum_{i=1}^n x_i,$$

$$\hat{\beta} = \left[\sum_{i=1}^r (x_i - \bar{x}_2)^2 \right]^{-1} \sum_{i=1}^r (x_i - \bar{x}_2) (y_i - \bar{y}_2)$$

and the second phase units are indexed from one to r . It is well known (e.g., Cochran 1977, equation 12.51) that the variance of the regression estimator is, approximately,

$$V\{\hat{\mu}_y\} = [n^{-1}\rho^2 + r^{-1}(1-\rho^2)]\sigma_y^2, \quad (2)$$

where ρ is the population correlation between y and x and σ_y^2 is the population variance of y . An estimator of the variance is, by classical regression theory,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_I)^2 + r^{-1}(r-2)^{-1} \sum_{i=1}^r (y_i - \hat{y}_i)^2 \quad (3)$$

where $\hat{y}_i = \bar{y}_2 + (x_i - \bar{x}_2)\hat{\beta}$ for $i = 1, 2, \dots, n$, and $\bar{y}_I = n^{-1} \sum_{i=1}^n \hat{y}_i$. Observe that \bar{y}_I is an alternative way of writing $\hat{\mu}_y$ in (1).

Let

$$c_r = [n(n-1)r^{-1}(r-2)^{-1}]^{1/2} \quad (4)$$

and

$$y_i^* = \begin{cases} \hat{y}_i, & i = r+1, r+2, \dots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i), & i = 1, 2, \dots, r. \end{cases} \quad (5)$$

Then,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (y_i^* - \bar{y}_I)^2 \quad (6)$$

where \bar{y}_I is the mean of the y_i^* , as well as the mean of the \hat{y}_i , because the sum of $y_i - \hat{y}_i$ is zero. Equation (6) is the operational form of the suggested estimator. The variance estimation data set contains the pseudo observation y_i^* .

To the extent that the model for imputation matches that of two-phase sampling, equation (6) is applicable to an imputed data set. For example, if we assume that missing data are missing at random and use regression to impute the missing value with \hat{y}_i , then equation (6) is immediately applicable. Of course, regression imputation or two-phase sampling can use a vector x .

3. EXTENSIONS TO RANDOM IMPUTATION

A moderate extension of the method described in section 2 enables us to estimate the variance of a sample mean using random imputation. In fact, alternative approaches are possible.

As one approach, assume that the imputation model is the regression model

$$y_i = \mathbf{x}_i \beta + e_i \quad (7)$$

where the first element of every \mathbf{x}_i is equal to 1 and the e_i are uncorrelated $(0, \sigma_e^2)$ random variables.

Assume the model is estimated and that the imputed values are

$$\hat{y}_i = \hat{y}_i + \hat{e}_i, \quad i = r+1, r+2, \dots, n \quad (8)$$

where $\hat{y}_i = \mathbf{x}_i \hat{\beta}$ with $\hat{\beta} = (\sum_{i=1}^r \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_{i=1}^r \mathbf{x}_i' y_i$ and \hat{e}_i is chosen at random from the set $\hat{\mathbf{e}}_r = \{\hat{e}_i = y_i - \hat{y}_i; i = 1, 2, \dots, r\}$. The estimator of the mean of y is

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^n \hat{y}_i \quad (9)$$

where $\hat{y}_i = y_i$ if $i = 1, 2, \dots, r$.

If the \hat{e}_i are chosen with replacement with equal probability from the set $\hat{\mathbf{e}}_r$, then the variance $\hat{\mu}_y$ is, approximately,

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}m)(1 - R^2)]\sigma_y^2, \quad (10)$$

where $m = n - r$ and R^2 is the squared multiple correlation coefficient between y and \mathbf{x} . The increase in variance due to using random imputation with \hat{e}_i , rather than using $\hat{e}_i = 0$, is $n^{-2}m(1 - R^2)\sigma_y^2$.

Therefore, an estimator of the variance of the imputed sample mean is given by (6) where the c_r of (4) is

$$c_I = [n(n-1)(r^{-1} + n^{-2}m)(r-p)^{-1}]^{1/2}, \quad (11)$$

and p is the dimension of β . We have

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_I)^2 + (r^{-1} + n^{-2}m)(r-p)^{-1} \sum_{i=1}^r (y_i - \hat{y}_i)^2 \quad (12)$$

where $\bar{y}_I = \sum_{i=1}^n \hat{y}_i$. The estimator of the variance using c_I of equation (11) is an estimator of the unconditional variance, the average over all possible imputed sample. Derivations of (10) and (12) are given in Appendix A.

To consider an alternative variance estimation approach, we assume that a random selection procedure is used for imputation but place no restriction on the procedure, other than that the probabilities of selection are inversely proportional to the probability that the y -value responds. In addition, we record the number of times an \hat{e} value is used as a donor in the imputation.

Let

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, \dots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i) & i = 1, 2, \dots, r \end{cases} \quad (13)$$

with

$$c_r = [n^{-1}(n-1)r(r-p)^{-1}]^{1/2}(1 + d_i) \quad (14)$$

where d_i is the number of times \hat{e}_i is used as a donor. The term $[n^{-1}(n-1)r(r-p)^{-1}]^{1/2}$ is used to adjust for the effect of estimating p regression parameters. Then, the variance estimator (6) can be written as

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_I)^2 + n^{-2}r(r-p)^{-1} \sum_{i=1}^r (1 + d_i)^2 (y_i - \hat{y}_i)^2. \quad (15)$$

If the imputation method is simple random sampling with replacement, then, conditional on the sample and the respondents,

$$E_I\{(1 + d_i)^2\} = \left(\frac{n}{r}\right)^2 + \frac{m}{r} \left(1 - \frac{1}{r}\right) \quad (16)$$

where the notation I is used here to denote the expectation with respect to the imputation mechanism generated by random imputation. The equality in (16) establishes the equivalence of (12) to (15) under with-replacement selection. It is shown in Appendix B that $\hat{V}\{\hat{\mu}_y\}$ in (15) is also a valid estimator when donors are selected without replacement. Since the proposed variance estimation method is the conditional variance given the realized imputed sample, it has wide applicability.

4. COMPLEX SAMPLING DESIGNS

4.1 Deterministic Imputation

The suggested method is applicable to complex sampling designs as well as to simple random sampling. Assume that the full sample estimator of the mean of y can be written as $\bar{y} = \sum_{i=1}^n w_i y_i$, where w_i is the sampling weight of unit i in the sample. Assume that $\sum_{i=1}^n w_i = 1$.

If the first r elements are observed and the remaining $n - r$ elements are missing, then the estimator of the mean of y under regression imputation is

$$\bar{y}_I = \sum_{i=1}^r w_i y_i + \sum_{i=r+1}^n w_i \hat{y}_i \quad (17)$$

where

$$\hat{y}_i = \mathbf{x}_i' \hat{\beta},$$

$$\hat{\beta} = \left[\sum_{i=1}^r w_i^* \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \sum_{i=1}^r w_i^* \mathbf{x}_i' y_i.$$

Here w_i^* is the sampling weight of unit i in the second-phase sample and is defined by

$$w_i^* = [\text{Pr}(i \text{ is in the second phase sample} \mid i \text{ is in the first phase sample})]^{-1} w_i.$$

Also, $\sum_{i=1}^r w_i^* = 1$. If we assume that the second phase sample is a random sample of size r from the n first phase sample, then $w_i^* = nr^{-1}w_i$. Under certain conditions we can write the estimator in (17) as

$$\bar{y}_I = \sum_{i=1}^n w_i \hat{y}_i. \quad (18)$$

The representation (18) holds if $(w_i^*)^{-1} w_i$ is in the column space of the matrix $\mathbf{X} = (\mathbf{x}_1', \dots, \mathbf{x}_r')$ because then we have $\sum_{i=1}^r w_i (y_i - \hat{y}_i) = 0$ from $\sum_{i=1}^r w_i^* \mathbf{x}_i' (y_i - \hat{y}_i) = 0$.

We assume a sequence of samples and finite populations such as that described in Fuller (1998). Define $\bar{\mathbf{x}}_1 = \sum_{i=1}^n w_i \mathbf{x}_i$ and $(\bar{\mathbf{x}}_2, \bar{y}_2) = \sum_{i=1}^n w_i^* (\mathbf{x}_i, y_i)$. We also adopt the same assumptions as in Fuller (1998). That is

$$E(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2) = (\mu_x, \mu_x, \mu_y), \quad (19)$$

and

$$V\{(\hat{\beta} - \beta)', \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2\} = O(n^{-1}), \quad (20)$$

where $(\mu_x, \mu_y) = N^{-1} \sum_{i=1}^N (\mathbf{x}_i, y_i)$ and $\beta = (\sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_{i=1}^N \mathbf{x}_i' y_i$.

For $i = 1, 2, \dots, N$, define

$$a_i = \begin{cases} 1 & \text{if unit } i \text{ responds when sampled} \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{a} = (a_1, a_2, \dots, a_N)$. The extended definition of a_i is discussed by Fay (1991) and used in Shao and Steel (1999). Now, let

$$\bar{y}_{II} = \sum_{i=1}^n w_i \tilde{y}_i^* \quad (21)$$

where

$$\tilde{y}_i^* = \tilde{y}_i + a_i w_i^{-1} w_i^* (y_i - \tilde{y}_i) \quad (22)$$

with $\tilde{y}_i = \mathbf{x}_i \beta$. Then, we have $\bar{y}_I = \bar{y}_H + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. By (19) and (20), we have $\bar{y}_I = \bar{y}_H + O_p(n^{-1})$ and $V(\bar{y}_I - \bar{y}_N) = V(\bar{y}_H - \bar{y}_N) + o(n^{-1})$. Now,

$$V(\bar{y}_H - \bar{y}_N) = V[E(\bar{y}_H - \bar{y}_N | \mathbf{a})] + E[V(\bar{y}_H - \bar{y}_N | \mathbf{a})]. \quad (23)$$

The first term on the right side of (23) is 0 because $E(\bar{y}_H - \bar{y}_N | \mathbf{a}) = 0$ under model (7). To estimate the second term in (23), note that conditional on \mathbf{a} , \bar{y}_H is a linear estimator. Hence, the standard variance estimation method applied to the pseudo data set $\tilde{\mathbf{Y}}^* \equiv \{\tilde{y}_i^*; i = 1, 2, \dots, n\}$ will unbiasedly estimate the variance of $\bar{y}_H = \sum_{i=1}^n w_i \tilde{y}_i^*$. Since the set $\tilde{\mathbf{Y}}^*$ is not observable, we can use the set $\mathbf{Y}^* \equiv \{y_i^*; i = 1, 2, \dots, n\}$, where

$$y_i^* = \hat{y}_i + a_i w_i^{-1} w_i^* (y_i - \hat{y}_i) \quad (24)$$

to get a consistent variance estimator.

To illustrate that the set \mathbf{Y}^* can be used to approximate the variance estimator, assume that the full sample variance estimator of \bar{y} can be written as

$$\hat{V} = \sum_{i=1}^L c_i (\bar{y}^{(i)} - \bar{y})^2$$

where L is the number of replications, c_i is the i -th replication factor, and $\bar{y}^{(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j$ is the i -th replicate of \bar{y} . The term $M_j^{(i)}$ is the replication multiplier applied to the weight of unit j at the i -th replication. For example, under simple random sampling, the jackknife multiplier is

$$M_j^{(i)} = \begin{cases} (n-1)^{-1} n & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Assume that the replicate variance estimator \hat{V} is applied to the set \mathbf{Y}^* to get

$$\hat{V}^* = \sum_{i=1}^L c_i (\bar{y}_I^{*(i)} - \bar{y}_I)^2$$

where $\bar{y}_I^{*(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j^*$ with y_j^* being defined in (24). Then, we have

$$\bar{y}_I^{*(i)} - \bar{y}_I = \bar{y}_H^{*(i)} - \bar{y}_H + (\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)(\hat{\beta} - \beta) \quad (25)$$

where

$$(\bar{\mathbf{x}}_1^{(i)}, \bar{\mathbf{x}}_2^{(i)}) = \sum_{j=1}^n w_j M_j^{(i)} (\mathbf{x}_j, a_j w_j^{-1} w_j^* \mathbf{x}_j).$$

It is shown in Appendix C that

$$\hat{V}^* = \sum_{i=1}^L c_i (\bar{y}_H^{*(i)} - \bar{y}_H)^2 + o_p(n^{-1}). \quad (26)$$

Therefore, the standard jackknife variance estimator applied to the pseudo data set \mathbf{Y}^* can be used to approximate the

standard jackknife variance estimator applied to the pseudo data set $\tilde{\mathbf{Y}}^*$.

4.2 Random Imputation

The arguments for variance estimation with random imputation are quite similar to those for deterministic imputation described in the previous subsection. First, define the imputation indicator function

$$d_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is used as donor for unit } j \\ 0 & \text{otherwise.} \end{cases} \quad (27)$$

Then, the estimator of the mean of y using random imputation is

$$\bar{y}_I = \sum_{i=1}^n w_i y_i^* \quad (28)$$

where

$$\bar{y}_i^* = \hat{y}_i + a_i (1 + d_i) (y_i - \hat{y}_i) \quad (29)$$

and

$$d_i = \sum_{j=1}^n (1 - a_j) d_{ij} w_i^{-1} w_j. \quad (30)$$

If the original sample weights are the same, then d_i is the number of times that unit i is used as a donor. We assume that

$$E[a_i (1 + d_i) | F_1] = 1 \quad (31)$$

where $F_1 = \{(i, \mathbf{x}_i, y_i); i = 1, 2, \dots, n\}$. The expectation in (31) is with respect to the joint distribution of the response mechanism and the imputation mechanism. Then, we have

$$E(\bar{y}_I | F_1) \doteq \bar{y}.$$

If we assume equal response probability, then, by (31), the probability of selection of donors should be proportional to the weights. This is the Rao and Shao (1992) setup for random imputation.

Now, let

$$\bar{y}_H = \sum_{i=1}^n w_i [\bar{y}_i + a_i (1 + d_i) (y_i - \bar{y}_i)] \quad (32)$$

where $\bar{y}_i = \mathbf{x}_i \beta$. Then, we also have $\bar{y}_I = (\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1)(\hat{\beta} - \beta) \bar{y}_H$ where $\bar{\mathbf{x}}_d = \sum_{i=1}^n w_i a_i (1 + d_i) \mathbf{x}_i$. By the assumption (31), we have $E(\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 | F_1) = 0$. Under mild conditions, $\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 = O_p(n^{-1/2})$ and $\bar{y}_I = \bar{y}_H + O_p(n^{-1})$. Now,

$$V(\bar{y}_I - \bar{y}_N) = V[E(\bar{y}_I - \bar{y}_N | \mathbf{a}, \mathbf{d})] + E[V(\bar{y}_I - \bar{y}_N | \mathbf{a}, \mathbf{d})]$$

where $\mathbf{d} = (d_1, d_2, \dots, d_n)$. Conditional on \mathbf{a} and \mathbf{d} , the estimator \bar{y}_H is a linear estimator. Hence, the pseudo data

$$y_i^* = \hat{y}_i + a_i (1 + d_i) (y_i - \hat{y}_i) \quad (33)$$

can be used to estimate the variance of \bar{y}_I .

5. COMPARISONS WITH ADJUSTED JACKKNIFE METHOD

Rao and Sitter (1995) proposed an adjusted jackknife variance estimator for the ratio imputation problem. Under the setup described in section 4, the ratio imputed estimator of μ_y is

$$\hat{\mu}_I = \sum_{i=1}^n w_i [a_i y_i + (1 - a_i) \hat{y}_i]$$

with $\hat{y}_i = x_i \hat{R}$ and $\hat{R} = (\sum_{i=1}^n w_i a_i x_i)^{-1} \sum_{i=1}^n w_i a_i y_i$. The Rao and Sitter (1995) variance estimator is

$$V_a = \sum_{i=1}^L c_i (\hat{\mu}_I^{(i)} - \hat{\mu}_I)^2, \quad (34)$$

where the adjusted jackknife replicate at the i -th replication is

$$\hat{\mu}_I^{(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j^{*(i)} \quad (35)$$

where

$$y_j^{*(i)} = \begin{cases} x_j \hat{R}^{(i)} & \text{if } a_i = 1 \\ x_j \hat{R} & \text{if } a_i = 0 \end{cases} \quad (36)$$

with $\hat{R}^{(i)} = (\sum_{j=1}^n w_j M_j^{(i)} a_j x_j)^{-1} \sum_{j=1}^n w_j M_j^{(i)} a_j y_j$. The adjusted values (36) in the Rao and Sitter (1995) method can also be regarded as pseudo data for variance estimation. Note that the calculation of the pseudo data (36) requires recalculation of $\hat{R}^{(i)}$ for each i with $a_i = 1$.

We modify the calculation of the pseudo values y_i^* in (5) to

$$y_i^* = \begin{cases} \hat{y}_i & \text{if } a_i = 0 \\ \hat{y}_i + c_r \left(\frac{\bar{x}_1}{\bar{x}_2} \right) (y_i - \hat{y}_i) & \text{if } a_i = 1, \end{cases} \quad (37)$$

where $\bar{x}_2 = \sum_{i=1}^n w_i r^{-1} n a_i x_i$, $\bar{x}_1 = n^{-1} \sum_{i=1}^n w_i x_i$ and $c_r = r^{-1} n$. The term (\bar{x}_1/\bar{x}_2) is inserted to improve the conditional properties of V_a given the first phase sample. The resulting variance estimator is approximately equivalent to the adjusted jackknife variance estimator (34). To see this, note that the adjusted values (35) can be written in the form

$$\hat{\mu}_I^{(i)} = \left(\sum_{j=1}^n w_j M_j^{(i)} x_j \right) \frac{\sum_{j=1}^n w_j M_j^{(i)} a_j y_j}{\sum_{j=1}^n w_j M_j^{(i)} a_j x_j} =: \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}},$$

where $A=:B$ denotes that we define B to be A . Also, define $\hat{Z} = \sum_{j=1}^n w_j x_j$, $\hat{S} = \sum_{j=1}^n w_j a_j y_j$, and $\hat{T} = \sum_{j=1}^n w_j a_j x_j$.

Then by the first order Taylor expansion,

$$\begin{aligned} \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}} &\doteq \hat{Z} \frac{\hat{S}}{\hat{T}} + (\hat{Z}^{(i)} - \hat{Z}) \frac{\hat{S}}{\hat{T}} \\ &\quad + (\hat{S}^{(i)} - \hat{S}) \frac{\hat{Z}}{\hat{T}} - (\hat{T}^{(i)} - \hat{T}) \frac{\hat{Z}\hat{S}}{\hat{T}^2} \\ &= \left[\hat{Z}^{(i)} \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} \left(\hat{S}^{(i)} - \hat{T}^{(i)} \frac{\hat{S}}{\hat{T}} \right) \right]. \end{aligned} \quad (38)$$

Note that the right side of (38) is exactly equal to

$$\sum_{j=1}^n w_j M_j^{(i)} \left[\frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_j \left(y_j - \frac{\hat{S}}{\hat{T}} \right) \right].$$

Thus, the pseudo data for variance estimation can be written as

$$y_i^* = \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_i \left(y_i - \frac{\hat{S}}{\hat{T}} \right),$$

which reduces to (37). Hence, the proposed method is exactly a first order Taylor linearization of the Rao and Sitter method in the case of ratio imputation. Therefore, we can expect our proposed method to have the same asymptotic properties as the Rao and Sitter method up to the order of n^{-1} .

The variance estimation method using the pseudo data set calculated by (37) is easy to implement because we can directly use existing software, which is more difficult with the Rao and Shao (1992) or Rao and Sitter (1995) method. Furthermore, if we calculate the pseudo data by (13), then the data set works for without-replacement hot deck imputation as well as for with-replacement hot deck imputation.

6. A SIMULATION STUDY

The preceding theory was tested in a simulation study using an artificial, finite population, from which repeated samples were drawn. The population has $L = 32$ strata, N_h clusters in stratum h , and 20 ultimate units in each cluster. The values of the population parameters were chosen to correspond to real populations encountered in the U.S. National Assessment of Educational Progress Study (Hansen and Tepping 1985) and are listed in Table 1. The finite population units are

$$y_{hij} = y_{hi} + e_{hij},$$

where

$$y_{hi} \stackrel{\text{iid}}{\sim} N(\mu_h, \sigma_h^2), \quad h = 1, 2, \dots, L, i = 1, 2, \dots, N_h,$$

and

$$e_{hij} \stackrel{iid}{\sim} N\left(0, \frac{1-\rho}{\rho} \sigma_h^2\right), j = 1, 2, \dots, 20.$$

Shao, Chen and Chen (1998) also used the same population in their simulation study. The value of the intra-cluster correlation ρ considered in the simulation is $\rho = 0.3$. Simulations with other values of ρ produced similar results and are not listed here for brevity.

Table 1
Parameters of the Finite Population for Simulation

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	100.0	20.0	2	16	95.0	19.0
3	20	90.0	18.0	4	25	98.0	19.6
5	25	93.0	18.6	6	25	98.0	19.6
7	25	96.0	19.2	8	28	94.0	18.8
9	28	92.0	18.4	10	28	96.0	19.2
11	31	94.0	18.8	12	31	92.0	18.4
13	31	90.0	18.0	14	31	96.0	19.2
15	31	94.0	18.8	16	31	92.0	18.4
17	31	90.0	18.0	18	31	88.0	17.6
19	31	86.0	17.2	20	34	84.0	16.8
21	34	82.0	16.4	22	34	80.0	16.0
23	34	90.0	18.0	24	37	85.0	17.0
25	37	80.0	16.0	26	37	90.0	18.0
27	37	85.0	17.0	28	39	80.0	16.0
29	39	75.0	15.0	30	42	75.0	15.0
31	42	75.0	15.0	32	42	75.0	15.0

We consider a stratified cluster sampling design, where $n_h = 2$ clusters are selected with replacement from stratum h with equal probability and all of the ultimate units in the selected clusters are in the sample. The sampling fraction is 6.4%. For each sampled unit y_{hij} , a response indicator variable a_{hij} is generated from

$$a_{hij} \stackrel{iid}{\sim} \text{Bernoulli}(p),$$

and that a_{hij} is independent of y_{hij} . The value of p considered in the simulation are $p = 0.9, 0.8, 0.7, 0.6$, and 0.5.

A set of 5,000 samples were selected using the same sampling design. In each of the selected samples, three imputation methods are considered;

- [M1] With-replacement weighted hot deck imputation considered by Rao and Shao (1992), where a missing value is imputed by a value randomly selected from the respondents with replacement with probability proportional to the survey weights.
- [M2] Without-replacement weighted hot deck imputation, which is the same as [M1] except that the selection was performed using a without-replacement sample. The without-replacement selection of donors is carried out systematically using the method described by Hansen, Hurwitz, and Madow (1953, page 343) from the respondents sorted by random order.

- [M3] Overall mean imputation, where the weighted mean of the respondents in the sample is imputed.

Hence, all the imputation methods use a single imputation cell that collapses all the strata.

In each imputed data set we computed three variance estimators \hat{V}_n , naive variance estimator treating the imputed data as if it were observed data, \hat{V}_a , the adjusted jackknife variance estimator of Rao and Shao (1992) for [M1] and [M2] and of Rao and Sitter (1995) for [M3], and \hat{V}^* , the jackknife variance estimator based on the pseudo data. The pseudo data set is constructed by (29) for [M1] and [M2] and by (24) for [M3]. The complete sample variance estimator used a standard jackknife for stratified cluster sampling, in which a cluster is deleted for each replication. Note that the standard jackknife is a consistent estimator of the variance under the model with nonzero intracluster correlation. Thus, the standard jackknife method based on the pseudo data can be applicable to the data set considered. The point estimators of the population mean are unbiased under the three different imputation schemes and are not listed here.

Table 2 presents the relative bias of the three variance estimators, the standard error of the relative bias of the variance estimators, and the sample correlation coefficient between the Rao's adjusted jackknife variance estimator and the new variance estimator based on the 5,000 samples. The relative bias of \hat{V} as an estimator of the variance of \bar{y}_I is calculated by $[\text{Var}_B(\bar{y}_I)]^{-1} [E_B(\hat{V}) - \text{Var}_B(\bar{y}_I)]$, where the subscript B denotes the distribution generated by the Monte Carlo simulation. The correlation coefficients of the two variance estimators are computed to give a measure the relative linearity behavior of the two variance estimators.

Table 2
Relative Bias of the Variance Estimator, Standard Error of the Relative Bias, and Sample Correlation Coefficient Between the Rao's Variance Estimator and the New Variance Estimator Based on 5,000 Samples

Response Rate (p)	Imputation Method	Rel. Bias $\times 100$ (S.E. $\times 100$)			Corr. Coeff. r
		Naive	Rao	New	
0.9	M1	-17.40 (2.02)	1.61 (2.03)	1.70 (2.04)	0.967
	M2	-17.50 (2.00)	1.41 (2.01)	0.81 (2.03)	0.974
	M3	-18.03 (2.03)	1.16 (2.05)	1.15 (2.04)	1.000
0.8	M1	-34.45 (2.01)	0.65 (2.03)	0.49 (2.05)	0.939
	M2	-32.89 (2.01)	2.49 (2.04)	0.19 (2.03)	0.947
	M3	-34.96 (2.01)	1.59 (2.03)	1.59 (2.03)	1.000
0.7	M1	-48.96 (2.01)	0.21 (1.99)	0.41 (2.04)	0.912
	M2	-44.76 (2.02)	5.31 (2.05)	0.76 (2.05)	0.920
	M3	-50.21 (2.02)	1.53 (2.05)	1.52 (2.04)	1.000
0.6	M1	-59.80 (2.02)	1.58 (2.05)	1.27 (2.06)	0.892
	M2	-54.86 (2.03)	7.10 (2.07)	-0.75 (2.07)	0.899
	M3	-64.11 (2.00)	-0.35 (2.04)	-0.35 (2.01)	1.000
0.5	M1	-69.75 (1.99)	0.84 (2.03)	1.12 (2.03)	0.873
	M2	-59.90 (2.01)	15.07 (2.07)	2.27 (2.06)	0.872
	M3	-74.44 (1.97)	1.99 (2.00)	1.98 (2.00)	1.000

Table 2 supports our theory in the following ways.

1. As is well known, the naive variance estimator seriously underestimates the true variance. The adjusted jackknife variance estimator performs well for [M1] and [M3], but not for [M2]. The theory for the adjusted jackknife method assumes that hot deck imputations are done using the with-replacement selection which is not used in [M2]. As the response rate decreases in Table 2, the relative bias of the adjusted jackknife becomes larger.
2. The new method based on the pseudo data performs well even for the without-replacement imputation [M2]. As was discussed at the end of section 3, a single formula (29) can be used as the pseudo data for a large class of imputation methods.
3. As is observed in the correlation coefficients, the behaviors of the adjusted jackknife variance estimator and the proposed variance estimator are very similar for mean imputation [M3]. This is because the two variance estimators are asymptotically equivalent, as discussed in section 5.

7. CONCLUDING REMARKS

We have described methods of making pseudo data to be used for variance estimation. Generally speaking, the pseudo data can be described as

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, \dots, n \\ \hat{y}_i + c_i g_i (y_i - \hat{y}_i) & i = 1, 2, \dots, r, \end{cases} \quad (39)$$

where \hat{y}_i is the predicted value of y_i under the model used for imputation. If $c_i g_i = 1$, then the variance estimator treats the imputed values as observations. A suitable choice of $c_i g_i > 1$ leads to a consistent variance estimator. If the imputation method is deterministic and the respondents are regarded as a random sample from the original sample, then $c_i \doteq r^{-1}n > 1$. For a two-phase sampling with a complex design, $c_i = w_i^{-1}w_i^*$, where w_i is the sampling weight of the unit i for the first-phase sample and w_i^* is the sampling weight of the unit i for the second-phase sample.

The g_i in (39) is the adjustment made to improve the conditional properties given the auxiliary variable x . For ratio imputation,

$$g_i = (\bar{x}_2)^{-1} \bar{x}_1$$

where $\bar{x}_2 = \sum_{i=1}^n w_i^* x_i$ and $\bar{x}_1 = \sum_{i=1}^n w_i x_i$. For regression imputation with scalar x ,

$$g_i = 1 + (\bar{x}_1 - \bar{x}_2) \left\{ \sum_{k=1}^r w_k^* (x_k - \bar{x}_2)^2 \right\}^{-1} (x_i - \bar{x}_2).$$

In either case, we have

$$\sum_{i=1}^r w_i^* g_i x_i = \bar{x}_1.$$

While this paper was under review, Shao and Steel (1999) also provided similar methods in the case of deterministic imputation. Our method is more general in the sense that we also considered random imputation and introduced c_i term to improve finite sample properties.

ACKNOWLEDGEMENTS

The author thanks his thesis adviser Wayne A. Fuller for valuable discussions. The author also thanks Pamela Abbitt, F. Jay Breidt, Lou Rizzo, Richard Valliant, and the referees for helpful comments, which greatly improved the paper. Most of this work was done while the author was a graduate student at Iowa State University and was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Service and Iowa State University and by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of Census.

APPENDIX

A. Proof of Equation (10) and (12)

The estimator $\hat{\mu}_y$ in (9) can be written as

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^r (1 + d_i) \hat{e}_i \quad (A.1)$$

where d_i is the number of times that unit i is used as a donor. Under the equal probability and with-replacement imputation mechanism, we have

$$E_I(d_i) = r^{-1}m$$

and

$$\text{Cov}_I(d_i, d_j) = \begin{cases} r^{-1}m(1 - r^{-1}) & \text{if } i = j \\ -r^{-2}m & \text{if } i \neq j \end{cases}$$

where the subscript I denotes the variation due to the imputation mechanism. It follows that $E_I(\hat{\mu}_y) = n^{-1} \sum_{i=1}^n \hat{y}_i$ and $V_I(\hat{\mu}_y) = n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2$. Hence,

$$V(\hat{\mu}_y) \doteq V \left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) + E \left(n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2 \right) \quad (A.2)$$

Now, by an similar argument similar to the one leading to (2), we have

$$\text{Var} \left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) = [n^{-1} R^2 + r^{-1} (1 - R^2)] \sigma_y^2. \quad (\text{A.3})$$

Since $\hat{y}_i - \bar{y}_t = (\mathbf{x}_i - \bar{\mathbf{x}}_t) \beta + o_p(1)$, we apply classical regression theory to get

$$E \left[(r-p)^{-1} \sum_{i=1}^r \hat{e}_i^2 \right] = (1 - R^2) \sigma_y^2, \quad (\text{A.4})$$

and

$$E \left[(n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_t)^2 \right] = R^2 \sigma_y^2. \quad (\text{A.5})$$

Therefore, (10) is proved and the estimator in (12) is consistent for the variance in (10).

B. Validity of (15) Under the Without-Replacement Imputation Mechanism

We assume that $m = kr + t$ where k and t are nonnegative integers and $t < r$. Let the estimator of the mean of y have the form (A.1). Let the imputation be performed such that t of the respondents are used $k+1$ times for imputation and $r-t$ units are used k times for imputation. The t of the respondents that are used $k+1$ times are chosen by simple random sampling without replacement. Then,

$$E_I(d_i) = k + r^{-1}t = r^{-1}m$$

and

$$\text{Cov}_I(d_i, d_j) = \begin{cases} r^{-1}t(1 - r^{-1}t) & \text{if } i = j \\ -r^{-2}t & \text{if } i \neq j. \end{cases}$$

So, by similar arguments as in the proof of (A.2), we have

$$V(\hat{\mu}_y) \doteq V(\bar{y}_t) + E \left[n^{-2} r^{-1} t \sum_{i=1}^r \hat{e}_i^2 \right]. \quad (\text{B.1})$$

Hence, using (A.3) and (A.4), we have

$$V(\hat{\mu}_y) = [n^{-1} R^2 + (r^{-1} + n^{-2}t)(1 - R^2)] \sigma_y^2. \quad (\text{B.2})$$

Now, conditional on the realized sample and the respondents, we have

$$E_I \left\{ (1 + d_i)^2 \right\} = \left(\frac{n}{r} \right)^2 + \frac{t}{r} \left(1 - \frac{t}{r} \right)$$

so that $\hat{V}\{\mu_y\}$ in (15) satisfies

$$\begin{aligned} E_I \left(\hat{V}\{\mu_y\} \right) &\doteq n^{-1} (n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y}_t)^2 \\ &\quad + [r^{-1} + n^{-2}t(1 - r^{-1}t)] \\ &\quad (r-p)^{-1} \sum_{i=1}^r (y_i - \hat{y}_i)^2. \end{aligned}$$

Therefore, using (A.4) and (A.5), we have the approximate unbiasedness of the $\hat{V}\{\mu_y\}$ under the without-replacement imputation mechanism.

C. Proof of Equation (26)

First, define $R_n^{(i)} = (\mathbf{x}_1^{(i)} - \bar{\mathbf{x}}_2) (\hat{\beta} - \beta)$ and $R_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) (\hat{\beta} - \beta)$. From the equality (25),

$$\hat{V}^* = \sum_{i=1}^L c_i \left(\bar{y}_I^{*(i)} - \bar{y}_I \right)^2 = A_n + B_n + 2C_n$$

where $A_n = \sum_{i=1}^L c_i (\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2$, $B_n = \sum_{i=1}^L c_i (R_n^{(i)} - R_n)^2$, and $C_n = \sum_{i=1}^L c_i (\bar{y}_{II}^{*(i)} - \bar{y}_{II}) (R_n^{(i)} - R_n)$. Hence, by the assumption (20), (26) follows because $A_n = O_p(n^{-1})$, $B_n = O_p(n^{-1})$, and $C_n = O_p(n^{-1})$. The last property comes from the Cauchy-Schwartz inequality, $C_n^2 \leq A_n B_n$.

REFERENCES

- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons.
- FAY, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual Research conference*, 429-440.
- FAY, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 227-232.
- FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- FULLER, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- HANSEN, M., HURWITZ, W.N. and MADOWS, W.G. (1953). *Sample Survey Methods and Theory*, Vol. I, New York: John Wiley and Sons.
- HANSEN, M., and TEPPING, B.J. (1985). Estimation for Variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.
- RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

- RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SÄRNDAL, C.-E. (1992). Methods for estimating the precision when imputation has been used. *Survey Methodology*, 18, 241-252.
- SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.
- SHAO, J., and STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. *Journal of the American Statistical Association*, 94, 254-265.

A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models

TRIVELLORE E. RAGHUNATHAN, JAMES M. LEPKOWSKI, JOHN VAN HOEWYK
and PETER SOLENBERGER¹

ABSTRACT

This article describes and evaluates a procedure for imputing missing values for a relatively complex data structure when the data are missing at random. The imputations are obtained by fitting a sequence of regression models and drawing values from the corresponding predictive distributions. The types of regression models used are linear, logistic, Poisson, generalized logit or a mixture of these depending on the type of variable being imputed. Two additional common features in the imputation process are incorporated: restriction to a relevant subpopulation for some variables and logical bounds or constraints for the imputed values. The restrictions involve subsetting the sample individuals that satisfy certain criteria while fitting the regression models. The bounds involve drawing values from a truncated predictive distribution. The development of this method was partly motivated by the analysis of two data sets which are used as illustrations. The sequential regression procedure is applied to perform multiple imputation analysis for the two applied problems. The sampling properties of inferences from multiply imputed data sets created using the sequential regression method are evaluated through simulated data sets.

KEY WORDS: Item nonresponse; Missing at random; Multiple imputation; Nonignorable missing mechanism; Regression; Sampling properties and simulations.

1. INTRODUCTION

Incomplete data is a pervasive problem faced by most applied researchers. Several methods have been, and continue to be, developed to draw inferences from data sets with missing values (Little and Rubin 1987). The multiple imputation framework suggested by Rubin (1978, 1987a, 1996) is an attractive option if a data set is to be used by multiple researchers with differing levels of statistical expertise. This approach involves imputing several plausible sets of missing values in the incomplete data set resulting in several completed data sets. Each completed data set is analyzed separately, say by fitting a particular regression model. The resulting inferences – point estimates and the covariance matrices – are then combined using the formula given in Rubin (1987a, Chap. 3) and refinements thereof (Li, Raghunathan and Rubin 1991; Li, Meng, Raghunathan and Rubin 1991; Meng and Rubin 1992; and Barnard 1995).

Imputation based approaches for handling missing data, in general, are quite useful in practice because once the missing values have been imputed, existing complete-data software can be used to analyze the data. Since software development for complete data analysis is keeping pace with the introduction of new statistical methods, applied researchers without knowledge of particular missing data techniques or resources to generate their own code for implementing new missing data procedures will be able to fit finely tuned substantive models for a specific problem at

hand. An added advantage of the multiple imputation approach is that by repeatedly applying the complete data software, one can obtain valid point and interval estimates under a fairly general set of conditions (Rubin 1987a). Several researchers (see, for example, the list of references in Rubin 1996) have applied this technique under a variety of settings and have demonstrated, through analysis of simulated and actual data sets, the appropriateness of this approach. Alternatives such as single imputation with an appropriate variance estimation procedure, for example, modified Jackknife Repeated Replication Technique (Rao and Shao 1992) also have this advantage. The imputation approach described in this paper can also be used to create single imputation with an alternative variance estimation procedure.

The development of imputation methods from varying perspectives has a long history (Madow, Nisselson, Olkin and Rubin 1983). A theoretically appealing framework for developing imputation methods is the Bayesian approach. This approach specifies an explicit model for variables with missing values, conditional on the fully observed variables and some unknown parameters, a prior distribution for the unknown parameters, and a model for the missing data mechanism, which does not need to be specified under an ignorable missing data mechanism (Rubin 1976). This explicit model then generates a posterior predictive distribution of the missing values conditional on the observed values. The imputations are drawn from this posterior predictive distribution. Several computer programs and

¹ Trivellore E. Raghunathan, James M. Lepkowski, John van Hoewyk and Peter Solenberger, University of Michigan, Institute for Social Research, Survey Methodology Program, P.O. Box 1248, Ann Arbor, MI 48106-1248, U.S.A.

algorithms are available for imputing missing values under multivariate normality (Rubin and Schafer 1990), the multivariate t distribution (Liu 1995), and several variations of the general location model (Schafer 1997; Raghunathan and Grizzle 1995; and Raghunathan and Siscovick 1996). The latter model can handle the joint distribution of categorical and continuous variables and was first proposed by Olkin and Tate (1961), and used by Little and Schluchter (1985) explicitly for missing data problems. An important property of these approaches is that they are fully conditional on all the observed information. Several simulation studies (for example, Raghunathan and Grizzle 1995) indicate that the inferences drawn from such imputed data have desirable sampling properties.

Survey data sets often consist of large numbers of variables which have a variety of distributional forms. Typically, such data sets have hundreds of variables, some continuous, others counts, many dichotomous or polytomous, and even some semi-continuous or limited dependent variables. Moreover, the distributions of the continuous variables alone may involve normal, lognormal, and other distributions. Postulating a full Bayesian model can be very difficult in this situation. Furthermore, survey data commonly have two additional features that make the modeling process even more complex. First, certain restrictions are imperative. For example, the variable "Number of Years Since Quit Smoking" is defined only for former smokers; hence, the imputation process for this variable should be restricted only to former smokers. Restrictions also arise due to skip patterns in the questionnaire. For example, certain questions about income from a second job are asked only when the respondent indicates that he/she has a second job. The imputation of such variables has to be handled in a hierarchical manner.

Second, there are certain logical or consistency bounds for the missing values that must be incorporated in the imputation process. Such interrelationships among the variables make the model specification difficult. For instance, "Years of Smoking" is restricted to current or past smokers and the imputed values must be less than $\text{Age} - x$ years, where x may be chosen based on certain other characteristics, such as evidence of smoking as a teen-ager. For a former smoker, x also includes years since smoking ceased. Another example of bounds is discussed in Heeringa, Little and Raghunathan (1997). They address imputation of bracketed response questions in which a respondent is unable or unwilling to provide an exact response (e.g., income and assets), but does define the bounds within which the imputed values must lie.

The goal of this paper is to propose and evaluate a general purpose multivariate imputation procedure that can handle a relatively complex data structure where explicit full multivariate models cannot be easily formulated but the imputed values for each individual are fully conditional on all the values observed for that individual. The approach is to consider imputation on a variable by variable basis but to

condition on all observed variables. The basic strategy creates imputations through a sequence of multiple regressions, varying the type of regression model by the type of variable being imputed. Covariates include all other variables observed or imputed for that individual. The imputations are defined as draws from the posterior predictive distribution specified by the regression model with a flat or non-informative prior distribution for the parameters in the regression model. The sequence of imputing missing values can be continued in a cyclical manner, each time overwriting previously drawn values, building interdependence among imputed values and exploiting the correlational structure among covariates. To generate multiple imputations, the same procedure can be applied with different random starting seeds or taking every P^{th} imputed set of values in the cycles mentioned above.

The variables in the data set are assumed to be of the following five types: (1) continuous, (2) binary, (3) categorical (polytomous with more than two categories), (4) counts and (5) mixed (a continuous variable with a non-zero probability mass at zero). Computationally, binary and categorical variables can be treated identically, but distinguishing them helps in conceptual understanding and in the description of the basic algorithm. We also assume that the population is essentially infinite, the sample is a simple random sample and the missing data mechanism is ignorable (Rubin 1976). The use of multiple imputation in a complex design setting has, as yet, not been fully investigated and is beyond the scope of the current paper.

In this paper we describe the sequential regression multivariate imputation (SRMI) approach in section 2 and evaluate two applications of the approach in sections 3 and 4. In the first application, it is difficult to postulate a joint multivariate distribution because of the complex systematic relationship between the variables and restrictions. In the second application, a general location model can be used to create multiple imputations (Olkin and Tate 1961; and Little and Schluchter 1985). Hence, we compare multiple imputation inferences resulting from the SRMI approach to those resulting from a joint multivariate model. The results of a simulation study investigating the sampling properties of imputed data inferences are presented in section 5, and a concluding discussion with directions for future research are given in section 6.

2. IMPUTATION METHOD

For a sample of size n , let X denote a $n \times p$ design or predictor matrix containing all the variables with no missing values. X consists of continuous, binary, count or mixed variables, and appropriate dummy variables representing categorical variables. In addition, X may also consist of a column of ones to model an intercept parameter, offset variables, and certain design variables. Let Y_1, Y_2, \dots, Y_k denote k variables with missing values, ordered, without

loss of generality, by the amount of missing values, from least to most. The pattern need not be monotone. (In a monotone pattern of missing data, Y_2 is observed only for a subset of subjects on whom Y_1 is observed, Y_3 is observed only for a subset of those on whom Y_2 is observed and so on.)

For model based imputations, the joint conditional density of Y_1, Y_2, \dots, Y_k given X can be factored as

$$f(Y_1, Y_2, \dots, Y_k | X, \theta_1, \theta_2, \dots, \theta_k) = f_1(Y_1 | X, \theta_1) f_2(Y_2 | X, Y_1, \theta_2) \dots f_k(Y_k | X, Y_1, Y_2, \dots, Y_{k-1}, \theta_k) \quad (1)$$

where $f_j, j=1, 2, \dots, k$ are the conditional density functions and θ_j is a vector of parameters in the conditional distribution (e.g., regression coefficients and dispersion parameters). In the sample survey context this can be viewed as a superpopulation model. We model each conditional density through an appropriate regression model with unknown parameters, θ_j , and draw from the corresponding predictive distribution of the missing values given the observed values. We assume that the prior distribution for the parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ is $\pi(\theta) \propto 1$ (diffuse relative to the likelihood). However, the method can easily be modified for specified proper prior distributions.

Each conditional regression is based on one of the following models:

1. A normal linear regression model on a suitable scale (for example, a Box-Cox power transformation may be used to achieve normality) if Y_j is continuous;
2. A logistic regression model if Y_j is binary;
3. A polytomous or generalized logit regression model if Y_j is categorical;
4. A Poisson loglinear model if Y_j is a count variable; and
5. A two-stage model where zero-non zero status is imputed using logistic regression, and conditional on non-zero status, a normal linear regression model is used to impute non-zero values, if Y_j is mixed.

Each imputation consists of c "rounds". Start round 1 by regressing the variable with the fewest number of missing values, Y_1 on X , imputing the missing values under the appropriate regression model. Assuming a flat prior for the regression coefficients, the imputations, for the missing values in Y_1 are the draws from the corresponding posterior predictive distribution (See Appendix A for a detailed discussion about drawing values for various regression models.) Then update X by appending Y_1 appropriately (for example, dummy variables, if it is categorical) and move on to the next variable, Y_2 , with the next fewest missing values. Repeat the imputation process using updated X as predictors until all the variables have been imputed. That is, Y_1 is regressed on $U = X$; Y_2 is regressed

on $U = (X, Y_1)$ where Y_1 has imputed values; Y_3 is regressed on $U = (X, Y_1, Y_2)$ where Y_1 and Y_2 have imputed values; and so on.

The imputation process is then repeated in rounds 2 through c , modifying the predictor set to include all Y variables except the one used as the dependent variable. Thus, regress Y_1 on X and Y_2, Y_3, \dots, Y_k ; regress Y_2 on X and Y_1, Y_3, \dots, Y_k ; and so on. Repeated cycles continue for a prespecified number of rounds, or until stable imputed values occur.

The procedure outlined above needs modification to incorporate restrictions and bounds. The restrictions are handled by fitting the models to an appropriate subset of individuals. For example, a Poisson regression model could be applied to impute any missing values for the variable "Number of Pregnancies." The imputation will be restricted to women in the sample. As a covariate, though, this variable may be treated differently when imputing subsequent variables. For instance, certain dummy variables may be created based on this variable, which are then appended to the matrix U before proceeding with the imputation of the next variable.

Consider another example, "Years Smoking Cigarettes," where the sample would be restricted to current or past smokers. If there is no evidence of smoking as a teenager, "Years Smoking Cigarettes" for a current smoker should satisfy the bound (0, Age - 18). If there is some indication of smoking as a teenager then the range may be restricted to, say (0, Age - 12). For a past smoker these ranges will be (0, Age - 18 - YRSQUIT) and (0, Age - 12 - YRSQUIT) respectively, where YRSQUIT is the years since the individual quit smoking. The appropriate regression model for this variable is a truncated version of the normal linear regression model (possibly on a transformed scale). The parameters, the regression coefficients and the residual variance need to be drawn from the corresponding posterior distributions. The imputations are then drawn from the corresponding truncated normal distribution conditional on the drawn value of the parameters.

It is difficult to draw values of parameters directly from their posterior distribution with truncated normal likelihoods. However, it can be easily computed for a given parameter value. The Sampling-Importance-Resampling (SIR) algorithm (Rubin 1987b, Raghunathan and Rubin 1988) can be used to draw from the actual posterior distribution. First, draw several trial parameter values from the posterior distribution without applying the bounds (untruncated normal linear regression model). Second, attach an importance ratio to each trial value, defined as the ratio of the actual posterior density with bounds to the trial density (the posterior density without bounds), both evaluated at the drawn value. Finally, resample a single parameter value with probability proportional to the importance ratios. This method requires careful monitoring of the distribution of importance ratios (Gelman, Carlin, Stern and Rubin 1995).

The bounds can also be applied to polytomous variables. For instance, suppose that a variable Y can take one of k values, but the observed data suggests that the missing value for a particular subject can either be j or l . The contribution to the likelihood from this subject corresponds to the conditional binomial distribution. The draws in the multinomial step (see Appendix A) are made from the conditional distribution for these two categories. That is, the imputed value is j with probabilities $s_{j\cdot} = P_{j\cdot}/(P_{j\cdot} + P_{l\cdot})$ and l with probability $1 - s_{j\cdot}$.

At the completion of the initial round of imputations, the first complete data set with no missing values is available. The factorization in Equation (1) defines a joint conditional distribution of Y_1, Y_2, \dots, Y_k , given X . If the pattern of missing data is monotone, the imputations in the first round are approximate draws from the joint posterior predictive density of the missing values given the observed values. Note that the draws from the logistic, polytomous, and count variables are from large sample approximations of the posterior density of the regression coefficients. It is possible to improve upon these approximations by using, for example, the SIR algorithm or another rejection algorithm in each subsequent round.

When the pattern of missing data is not monotone, one can develop a Gibbs sampling algorithm (Geman and Geman 1984; Gelfand and Smith 1990) corresponding to Model (1). For example, conditional on the drawn values of the parameters $\theta_2, \theta_3, \dots, \theta_k$ and the missing values drawn in the first round, the second round would draw values of θ_1 from the appropriate conditional posterior density which is proportional to the first term in Equation (1). Next draw the missing values in Y_1 conditional on this drawn value of the parameter θ_1 , all other observed or imputed values for that subject and other parameters, $\theta_2, \theta_3, \dots, \theta_k$ in the model. That is, the missing values in Y_j at round $(t+1)$ need to be drawn from the conditional density,

$$f_j^*(Y_j | \theta_1^{(t+1)}, Y_1^{(t+1)}, \dots, \theta_j^{(t+1)}, Y_{j+1}^{(t)}, \dots, \theta_k^{(t)}, Y_k^{(t)}, X), \quad (2)$$

computed based on the joint distribution in (1), where $Y_i^{(t)}$ is the imputed or observed values for variable Y_i at round t . Though this is conceptually possible, it is difficult even to compute this density in most practical settings with restrictions, bounds, and the types of variables being considered.

Our proposal is to draw missing values in Y_j at round $(t+1)$ from a predictive distribution corresponding to conditional density,

$$g_j(Y_j | Y_1^{(t+1)}, Y_2^{(t+1)}, \dots, Y_{j-1}^{(t+1)}, Y_{j+1}^{(t)}, \dots, Y_k^{(t)}, X, \phi_j), \quad (3)$$

where the conditional density g_j is specified by one of the regression models described earlier that depends upon the variable type for Y_j , and ϕ_j is the unknown regression parameters with diffuse prior. That is, the new imputed values for a variable are conditional on the previously imputed values of other variables, and the newly imputed values of variables that preceded the currently imputed variable. This proposal may be viewed as an approximation to an actual

Gibbs sampling where the conditional density (2) is approximated by the conditional density (3). Furthermore, this approximation can be improved by considering the SIR or some other rejection type algorithm if the conditional density in (2) can be computed up to a constant.

There are some other particular cases where this approximation is equivalent to drawing values from a posterior predictive distribution under a fully parametric model. For example, if all the variables are continuous and each conditional regression model is a normal linear regression model with constant variance, then the algorithm converges to a joint predictive distribution under a multivariate normal distribution with an improper prior for the mean and the covariance matrix.

It is theoretically possible that a sequence of draws based on densities in (3) may not converge to a stationary distribution, because these conditional densities may not be compatible with any multivariate joint conditional distribution of Y_1, Y_2, \dots, Y_k given X (Gelman and Speed 1993). Our empirical investigations using several practical data sets have not identified, so far, any such anomalies. In several large data sets, we find the conditional densities (2) and (3) to be quite similar. As discussed in sections 4 and 5, the draws from this approach are comparable to those based on an explicit Bayesian model.

3. EFFECT OF SMOKING ON PRIMARY CARDIAC ARREST

In our first illustration, the SRMI approach is applied to a case-control study examining the relationship between cigarette smoking and the incidence of primary cardiac arrest (Siscovick, Raghuathan, King, Weinmann, Wicklund, Albright, Bovbjerg, Arbogast, Kushi, Cobb, Copass, Psaty, Retzlaff, Childs and Knopp 1995). In this study it is difficult to formulate an explicit model which captures the full complexity of the data. The case subjects were all King County, Washington residents who had out-of-hospital primary cardiac arrests between 1988 and 1994. The case subjects were identified through a review of paramedic incident reports. Control subjects were selected by random digit dialing from King County and matched to case subjects on gender and age (within seven years). To be eligible, subjects (case and control) were required to be between 25 and 74 years of age, married, and free of clinically-diagnosed heart disease or some other life-threatening conditions such as cancer, liver disease, lung disease, or end-stage renal disease.

Because primary cardiac arrest has a case-fatality rate greater than 80%, the eligibility criterion of marriage was included so that information regarding risk factor exposure (*i.e.*, smoker status, years smoked) could be ascertained from surrogate respondents (*i.e.*, spouses). Among control and surviving cases subjects, both subject and surrogate were interviewed to gather exposure data. The control and

the surviving cases subjects were interviewed mainly to study the reliability of measurements from their surrogates. Among the variables considered in this paper, there were practically no differences in the measurements obtained from the subjects and their surrogates for control or case subjects.

Table 1 gives the means, standard deviations, and percent missing values for key variables by case-control status. The exposure variables are indicator variables for Former Smoker (X_1), Current Smoker (X_2) and Years Smoked (X_3). The confounding variables considered are Age, Body Mass Index (BMI) (BMI=Weight [in Kg]/Height²[in Meters]), and the binary variables Female and Education (High School Graduate). The substantive model of interest is the logistic regression model,

$$\begin{aligned} \log [\Pr(C = 1) / \Pr(C = 0)] = & \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_3 \\ & + \alpha_4 X_2 X_3 + \alpha_5 \text{Age} + \alpha_6 \text{BMI} \\ & + \alpha_7 \text{Female} + \alpha_8 \text{Education,} \end{aligned}$$

where C is an indicator of cardiac arrest. Preliminary investigations indicated that linear terms for Age and BMI, are appropriate.

Table 1
Means and Proportions (in %) for Key Variables and Percent Missing

Variable	Control (n=551)		Cases (n=347)	
	% Missing	Mean (SD)	% Missing	Mean (SD)
Age	0.0	58.4 (10.4)	0.0	59.4 (9.9)
BMI	8.2	25.8 (4.1)	2.6	26.4 (4.6)
Years Smoked	16.8	24.8 (14.7)	5.4	31.7 (13.8)
		Proportion		Proportion
Female	0.0	23.2	0.0	19.9
≥ High School	0.0	76.8	0.0	61.9
Smoking Status				
Never Smoked	0.0	47.2	0.0	27.3
Former Smoker	0.0	42.1	0.0	38.2
Current Smoker	0.0	10.7	0.0	34.5

There are no missing values for the variables Age, Female, Education, Smoking Status (X_1, X_2), and C . Thus, for purposes of imputation, define $X = (1, \text{Age, Female Education, } X_1, X_2, C)$. Log (BMI), having the fewest missing values, was regressed first on X through a normal linear regression model. Residual diagnostics indicated a log-transform improved the normality of residuals.

Next, Years Smoked was regressed on $U = (X, \log (\text{BMI}))$. For this variable the sample was restricted to current and former smokers. Moreover, imputed values for Years Smoked were bounded by AGE-18, unless a respondent reported that they smoked in school (SCHSMK), and then they were bounded by AGE-12. For former smokers, imputed values were also bounded by how long ago the respondent had quit smoking (YRSQUIT). Thus, imputed values for former smokers who did not

smoke in school were bounded by AGE-18-YRSQUIT, while imputed values for former smokers that did smoke in school were bounded by AGE-12-YRSQUIT. Some subjects (5%) had missing values on the two auxiliary items (SCHSMK, YRSQUIT) which were imputed prior to defining the upper bounds of Years Smoked. The inherent structure of this data set makes it difficult to develop explicitly a joint distribution of the variables with missing values conditional on the completed observed variables. SRMI is thus an appealing approach to handle for this type of data.

In imputing the missing values, we performed 1,000 rounds for each of 25 different starting random seeds resulting in $M = 25$ imputations. The logistic regression model was fit to each imputed data set to obtain maximum likelihood estimates of the regression coefficients and asymptotic covariance matrices.

We used the standard multiple imputation variance formula (Rubin 1987a, Chap. 3) to compute the multiply imputed estimate of the regression coefficients and the covariance matrix. Briefly, suppose that $\hat{\alpha}^{(l)}$ is the estimate of the vector of regression coefficients α in the logistic model, and $V^{(l)}$ its covariance matrix, based on imputed data set l . The multiply imputed estimate of α is

$$\hat{\alpha}_{MI} = \sum_{l=1}^M \hat{\alpha}^{(l)} / M$$

and its covariance matrix is

$$V_{MI} = \sum_{l=1}^M V^{(l)} / M + \frac{M+1}{M} B_M$$

where

$$B_M = \sum_{l=1}^M (\hat{\alpha}^{(l)} - \hat{\alpha}_{MI})(\hat{\alpha}^{(l)} - \hat{\alpha}_{MI})^t / (M - 1)$$

The number of imputations is larger than what is usually recommended. We performed 25 imputations with different random seeds to assess whether the Gibbs style rounds lead us to a region of the imputed values that is very different from the observed data. Graphical displays of the imputed and observed values indicated that none of the imputations in the 25,000 rounds were incompatible with the observed data distribution.

Table 2, the complete-case analysis, gives the point estimates and their standard errors based on subjects with all variables observed. A total of 103 subjects (11.5%) had missing values in one or more predictors. A complete-case analysis, which is generally valid only when the data are missing completely at random was performed after deleting these 103 subjects (See Column 2, Table 2). Logistic regression analyses with a missing data indicator as the dependent variable and a number of completely observed variables as predictors indicated that the data are not missing completely at random. One may expect, therefore, that the complete case estimates and standard errors are biased.

Table 2

Point Estimates (Standard Errors) of Logistic Regression Coefficients for Model of Primary Cardiac Arrest for Complete Cases, SRMI Methods 1* and 2**

Predictor Variables	Complete Case		SRMI			
	(n=795)		Method 1 (n=898)		Method 2 (n=898)	
	Estimate (SE)		Estimate (SE)		Estimate (SE)	
Intercept	-2.922	(0.791)	-2.610	(0.757)	-2.348	(0.627)
Age	0.015	(0.009)	0.015	(0.009)	0.014	(0.008)
Female	-0.007	(0.203)	-0.115	(0.189)	-0.119	(0.177)
Education	-0.448	(0.173)	-0.467	(0.166)	-0.444	(0.133)
BMI	0.056	(0.018)	0.049	(0.013)	0.055	(0.009)
Current Smoker	1.693	(0.569)	2.001	(0.543)	1.998	(0.448)
Former Smoker	0.003	(0.284)	-0.029	(0.262)	-0.011	(0.223)
Current Smoker × Yrs Smoked	-0.003	(0.015)	-0.008	(0.013)	-0.005	(0.011)
Former Smoker × Yrs Smoked	0.019	(0.009)	0.014	(0.009)	0.014	(0.009)

* Method 1 – Imputation restricted to model variables

** Method 2 – Imputation includes model and auxiliary variables

Table 2, SRMI Method 1, gives estimates and their standard errors for SRMI using only the variables in the substantive model. These estimates are quite similar to the complete-case analysis estimates. The multiple imputation standard errors are smaller due to additional subjects with imputed data. There are modest changes in the relationship between smoking and primary cardiac arrest. The complete-case analysis indicates a statistically significant relationship between years smoked and primary cardiac arrest for former smokers, while no such association is indicated in the analysis of multiply imputed data.

One of the advantages of the multiple imputation approach is that the imputation process can use additional variables not in the substantive analysis. Such situations arise when a common research database with many variables is used by different researchers, each using a subset of the variables. The imputation may be carried out for the entire database, where prediction for missing values in each variable borrows strength from all other variables in the data set. Such imputations have been shown to improve efficiency compared to those based only on variables in the particular substantive model (Raghunathan and Siscovick 1996).

Table 2, SRMI Method 2, provides multiple imputation estimates and their standard errors obtained when the entire data set was imputed using 50 additional variables. These included dietary indicators, physiological measures, socioeconomic status, and behavioural variables. The point estimates are modestly different for all the variables. The standard errors, though, are considerably smaller when compared to the multiple imputation approach using only variables in the substantive model (SRMI, Method 1). This is not surprising because many of the additional variables such as blood pressure, cholesterol counts, alcohol consumption, and physical activity were highly predictive of BMI and smoking related variables.

4. PARENTAL PSYCHOLOGICAL DISORDERS AND CHILD DEVELOPMENT

A second illustration examines the effects of parental psychological disorders on several measures of childhood development. Little and Schuchter (1985) analyzed the data using a general location model to obtain maximum likelihood estimates of the parameters of the joint distribution. This general location model was employed to create multiple imputations using Markov Chain Monte Carlo methods (Schafer 1997), producing fully Bayesian model-based multiply imputed data sets. We also created multiple imputations using the SRMI procedure.

The study data consists of 69 families with two children each. Each family was classified into one of the three risk categories: (1) Normal Risk – no parental psychiatric disorders; (2) Moderate Risk – one parent diagnosed with a psychiatric illness or a chronic physical illness; and (3) High Risk – one parent diagnosed with schizophrenia or an affective mental disorder. There are three primary dependent variables of interest: Y_{1c} , number of psychiatric symptoms (dichotomized as high/low) for child c ; Y_{2c} , the standardized reading scores for child c ; and Y_{3c} , the standardized verbal comprehension score for child c .

We consider three models in investigating the impact of parental psychological disorders on childhood development. The first is a mixed effects logistic regression model:

$$\text{logit}[\Pr(Y_{1ic} = 1)] = \beta_0 + \beta_1 U_{1i} + \beta_2 U_{2i} + \gamma_i$$

where $Y_{1ic} = 1$ if child c in family i is classified as having a high number of symptoms and 0 otherwise; $U_{1i} = 1$ if family i is classified as a moderate risk group and 0 otherwise; $U_{2i} = 1$ if family i is classified as a high risk group and 0 otherwise; and γ_i are random effects assumed to be identically and independently distributed normal random variables with mean 0 and variance ϕ_γ^2 . This

random effect accounts for intraclass correlation between the two children within the same family. With complete data, this model may be fit by maximizing the numerically integrated likelihood function of $(\beta_0, \beta_1, \beta_2, \phi_1^2)$ using the Newton-Raphson algorithm and the Gaussian quadrature method for the numerical integration of the likelihood function. These types of models can be easily fit with complete data, but are difficult to fit with missing data.

The second and third regression models relate the child's reading and verbal scores, respectively, to risk group after adjusting for the number of symptoms (Y_1). An investigation of the residuals after a few preliminary rounds of reading and verbal score imputations indicated a log scale was appropriate. Thus, denoting Y_{2ic} and Y_{3ic} as the logarithm of the reading and verbal scores, respectively, for child c in family i , we posited the following mixed effects regression model,

$$Y_{2ic} = \alpha_0 + \alpha_1 U_{1i} + \alpha_2 U_{2i} + \alpha_3 Y_{1ic} + \delta_i + \epsilon_{ic}.$$

where δ_i and ϵ_{ic} are mutually independent normal random variables with mean 0 and variances σ_δ^2 and σ_ϵ^2 respectively. Again, with no missing data in the covariates, the maximum likelihood estimates of the unknown parameters can be readily obtained using, for example, the PROC MIXED procedure in SAS.

There were no missing values in the classification of the risk groups, and thus we defined $X = (1, U_1, U_2)$. The variables with missing values, Y_{21}, Y_{22}, Y_{31} and Y_{32} were imputed using normal linear regression, and the missing values in Y_{11} and Y_{12} were imputed using logistic regression. We created $M=25$ SRMIs, repeating the process through 1,000 rounds and 25 different seeds. The SRMI multiply imputed data sets were analyzed and combined using the methods described earlier. To compare these results with the multiply imputed inferences when the imputations are draws from the posterior predictive distribution under the general location model we created 25 imputations under a fully Bayesian model using software developed by Schafer (1997). The point estimates and

standard errors for the three models using SRMI and Bayes multiple imputation approaches are presented in Table 3. There are no real meaningful differences between the SRMI estimates and standard errors and those resulting from the Bayesian imputation. Children of parents in the high risk group are approximately 7.8 [exp (2.048)] times more likely to have a high number of symptoms than children with parents in the normal group under the SRMI. The 95% confidence interval for this relative risk is (3.8, 16.0). For the moderate risk, group, the corresponding point and interval estimates are 3.7 and (1.8, 7.8). These estimates may be contrasted with those obtained based on the complete-case analysis (not shown): 7.4 (2.3, 24.2) for the high risk group, and 3.5 (1.0, 11.9) for the moderate risk group (data not shown). Though the point estimates of the relative risks are similar, the complete-case confidence intervals are wider because they are based only on 60% of the observations.

Based on the estimated regression coefficients in Table 3, one can infer, after adjusting, for the number of symptoms, that children in the moderate and high risk groups have lower reading scores, by about 11 points [exp (4.654)-exp(4.654-0.110)], when compared to the normal group. On the other hand, the complete-case analysis estimates a score of 16 points lower for children in the moderate risk group than their counterparts in the normal group, and children in the high risk group score about 19 points lower when compared to the normal group.

The SRMI analysis of verbal scores suggests that the children in the moderate and high risk groups score about 20 and 24 points lower, respectively, than their counterparts in the normal group. However, the complete-case analysis shows the moderate risk group scores lower by 36 points and the high risk group scores lower by about 39 points when compared to the normal group. Thus, the complete-case estimates of the effects of parental psychological disorders on the child's reading and verbal scores are quite different than those obtained by the analysis of the multiply imputed data. This is not surprising because the data on reading and verbal scores are not missing completely at

Table 3
Point Estimates (Standard Errors) of Regression Coefficients for Three Models of Child Development Under SRMI and Bayesian Imputation

Predictor Variables	Imp. Method	Dependent Variable					
		Symptoms		Reading Score		Verbal Score	
Intercept	SRMI	-0.678	(0.256)	4.654	(0.013)	4.873	(0.020)
	Bayes	-0.688	(0.257)	4.556	(0.013)	4.991	(0.021)
High Risk Group	SRMI	2.048	(0.356)	-0.109	(0.022)	-0.191	(0.032)
	Bayes	2.033	(0.350)	-0.108	(0.021)	-0.180	(0.033)
Moderate Risk Group	SRMI	1.289	(0.366)	-0.110	(0.022)	-0.162	(0.033)
	Bayes	1.300	(0.360)	-0.109	(0.023)	-0.167	(0.035)
Symptoms	SRMI	-	-	0.032	(0.022)	-0.083	(0.032)
	Bayes	-	-	0.031	(0.019)	-0.080	(0.030)

random and are related to the risk group as well as the number of symptoms of the child.

5. SIMULATION STUDY

The analyses described in sections 3 and 4 indicate that sensible results can be obtained by applying the SRMI approach to handling missing values. Nevertheless, it is difficult to conclude based on such case studies whether or not the approach will result in valid inferences in routine applications. A simulation study was designed to investigate the repeated sampling properties of inferences from imputed data sets created with the SRMI approach. Complete data sets were generated from hypothetical populations, and elements deleted under an ignorable missing data mechanism. The deleted values were imputed and differences in summary statistics based on the imputed data sets and the before deletion or full data sets were assessed.

More formally, the strategy:

- (1) generated a complete data set which did not agree perfectly with our multiple imputation strategy,
- (2) estimated selected regression parameters,
- (3) deleted certain values using an ignorable missing data mechanism,
- (4) used SRMI to multiply impute the missing values, and
- (5) obtained multiply imputed estimates for the regression parameters estimated in step 2.

The differences in the parameter are examined across several independent replications of this strategy.

A total of 2,500 complete data sets with three variables (U, Y_1, Y_2) and sample size 100 were generated using the following models:

1. $U \sim \text{Normal}(0, 1)$;
2. $Y_1 \sim \text{Gamma}$ with mean $\mu_1 = \exp(U-1)$ and variance $\mu_1^2/5$; and
3. $Y_2 \sim \text{Gamma}$ with mean $\mu_2 = \exp(-1 + 0.5U + 0.5Y_1)$ and variance $\mu_2^2/2$.

The model for Y_2 in step 3 is the primary regression model of interest with true regression coefficients $\beta_0 = -1$, $\beta_1 = \beta_2 = 0.5$, and dispersion parameter $\phi^2 = 0.5$. For the complete data this model can be fixed using statistical software packages such as GLIM or Splus.

The deletion or missing data mechanisms were as follows:

- (1) No missing values in U ;
- (2) the missing values in Y_1 depend on U through a logistic function $\text{logit}[\Pr(Y_1 \text{ is missing})] = 1.5 + U$; and
- (3) the missing values in Y_2 depend on U and Y_1 through a logistic function $\text{logit}[\Pr(Y_2 \text{ is missing})] = 1.5 - 0.5Y_1 - 0.5U$.

These missing data mechanisms generated 22% missing data in Y_1 and 29% missing data in Y_2 . The complete-case analysis would have only used 48% of the data.

Since SRMI allows us only to fit a normal linear regression model, the imputations were carried out as follows. Suppose that Y_1 has fewer missing values, and let $Z_1 = (Y_1^{\lambda_1} - 1)/\lambda_1$ be the Box-Cox transformation of the continuous variable. In the first round of imputations, assume that Z_1 has a normal distribution with mean $a_0 + a_1U$ and variance σ_1^2 , where λ_1 was estimated using the maximum likelihood approach, and that $Z_2 = (Y_2^{\lambda_2} - 1)/\lambda_2$ has a normal distribution with mean $b_0 + b_1U + b_2Z_1$ and variance σ_2^2 , where λ_2 was estimated using maximum likelihood. In the subsequent rounds, U and Z_2 are predictors for Z_1 , and U and Z_1 are predictors for Z_2 . The estimation of a power transformation using maximum likelihood was automated while fitting each regression model.

For each of the 2,500 simulated data sets with missing values, a total 250 rounds with $M=5$ different random starts were created using SRMI. For each replicate, the resulting $M=5$ imputed data sets and the full data set (before deletion) were analyzed by fitting the Gamma model for Y_2 using maximum likelihood. The multiple imputation estimate was constructed as the average of the five imputed data estimates. To assess the differences in the point estimates we computed the standardized difference between the SRMI and full data estimates,

$$\Delta(\beta) = \frac{100 \times \text{abs}(\text{SRMI estimate} - \text{Full Data Estimate})}{\text{SE}(\text{SRMI Estimate})}.$$

Table 4 gives the mean and standard deviation of $\Delta(\beta)$ for three regression coefficients β_0, β_1 , and β_2 in the model. The SRMI estimates are typically within 8% of the full standard units. The actual coverage and the average length of the 95% SRMI confidence intervals were computed for the regression coefficients using the t reference distribution described in Rubin (1987b). For each simulated data set and parameter, it was determined whether or not the true value (e.g., $\beta_1 = 0.5$) is contained within the corresponding interval. The proportion of intervals containing the true values were computed across the 2,500 replications and are provided in Table 4. For the full data sets, the actual coverage for β_1 , for example, was 94.9% and for SRMI it was 95.4. In addition the average length of the confidence intervals were also computed. The average width of the full data confidence interval for β_1 was 0.91 and for SRMI the average length was 1.22. That is, the SRMI data resulted in well calibrated intervals estimates.

The same simulation study was also used to compare the distributional properties of imputations from SRMI and a fully Bayesian method. For the model assumptions used to generate complete data, we developed a Markov Chain Monte-Carlo algorithm for drawing values from the actual posterior predictive distribution of the missing values given

the observed values. Each step of the draw used Metropolis-Hastings algorithm and required considerably more computational time than the SRMI method. Therefore, only the first 500 simulated data sets were used in this comparison. We computed two Kolmogrove-Smimoff (KS) statistics from each simulated data set: One comparing the imputations from the SRMI method and the actual hidden values and the other comparing the Bayesian imputations and the actual hidden values. There were no discernible differences in these two statistics across the 500 simulated data sets. A scatter plot of those 500 pairs of KS statistics showed a narrow scatter of points around a 45 degree line.

Table 4
Means and Standard Deviations for Standardized Differences Between SRMI Estimates and Full Data Estimates and Actual Coverage of Nominal 95% Confidence Intervals

Regression Coefficient	Std. Difference		Confidence Coverage	
	Mean	SD	SRMI	Full Data
β_0	8.2	2.0	96.1	95.4
β_1	8.8	1.7	95.4	94.9
β_2	8.0	2.2	95.3	94.7

6. DISCUSSION

We have described and evaluated a sequential regression multivariate imputation procedure that can be used to impute missing values in a variety of complex data structures involving many types of variables, restrictions, and bounds. This procedure should be useful when the specification of a joint distribution of all the variables with missing values is difficult. A real advantage of the approach is its flexibility in handling each variable on a case by case basis. For instance, to preserve all the bivariate correlations, all the main effect terms must be included as regressors, and to preserve, say, three factor interactions all two factor interactions must be included as regressors in the imputation model. Implementation of this procedure only requires a good random number generator and fitting routines for a variety of multiple regression routines. A SAS based application implementing this approach can be downloaded from a web site (www.isr.umich.edu/src/smp/ive).

In certain instances, one can modify the algorithm to reduce it to Gibbs sampling from the joint predictive distribution of the missing values given the observed values. However, the SRMI procedure will be more useful where an explicit model is difficult to formulate. In both the illustrations and the simulation, different random starts were used to monitor imputed values, an important aspect in many practical applications. This is a good practice when Gibbs sampling is used under an explicit Bayesian model (Gelman and Rubin 1992) and should be used when the sequential regression method discussed in this paper is used.

The simulation study described in section 5, though limited, is favorable as far as inferences based on the SRMI are concerned. The imputations from SRMI and Bayes model were comparable. The goal here, however, was to develop an imputation approach that is finely tuned on a variable by variable basis fully conditional on all the observed information, rather than an explicit joint multivariate distribution of all the variables. Furthermore, model sensitivity may be reduced by using a semiparametric regression model for each conditional regression. The Bayesian interpretation of the spline smoothing models (Silverman 1985) can be used to draw imputed values from the predictive distribution. Such modifications also deserve further investigation.

For some large data sets with many variables, the SRMI can be computationally intense. The algorithm can be modified to apply a variable selection method for each regression in each round. We compared the inferences with and without the variable selection on several large data sets such as the National Health Interview Survey and the National Medical Expenditure Survey using several hundred variables. The descriptive inferences as well as inferences based on linear and logistic regression models were very similar, still further detailed investigation is needed.

It is also possible to use the imputation approach discussed in this paper in conjunction with, for example, the Jackknife Repeated Replication (JRR) technique for variance estimation. Specifically, (1) re-impute, singly, the missing values in each jackknife replicate SRMI; (2) analyze the imputed replicate data set; and, finally, (3) combine the replicate estimates to obtain the point estimate and its covariance matrix. This approach is more computationally intensive than the multiple imputation approach. This integrated JRR imputation approach and several of its variations are currently under investigation.

Finally, it has been assumed that the data set arises from a simple random sample design. However, most surveys employ complex sample designs involving stratification, clustering, and weighting. Further work is needed to modify the sequential regression method to incorporate complex design features not reflected in the *X* variables in expression (1). However, even if the imputation process ignores the complex design features, the analysis of completed data should be design based. Though this does not provide valid design-based inferences, it maintains the robustness underlying the design-based analysis to a certain degree. The integrated JRR imputation approach discussed above may have more appealing design-based properties in a complex design setting.

ACKNOWLEDGEMENTS

The authors would like to thank the three referees for their careful reading of this article and their helpful suggestions. The research was partially supported by a NSF grant DMS-0803720.

APPENDIX: REGRESSION MODELS AND IMPUTATIONS

Dropping the subscript indexing of the variables for brevity, the necessary steps for imputing each type of variable are as follows:

Continuous variable: For Y (possibly transformed from the original scale for normality), a continuous variable, build a normal linear regression model, $Y = U\beta + e$, where U is the most recently updated predictor matrix, e has a multivariate normal distribution with mean zero and variance $\sigma^2 I$, and I is an identity matrix. Suppose that $\theta = (\beta, \log \sigma)$ has a uniform prior distribution over the appropriate dimensional real space. Fit this model based on the individuals for whom Y is observed.

Let $B = (U^T U)^{-1} U^T Y$ be the estimated regression coefficient, $SSE = (Y - UB)^T (Y - UB)$ be the residual sum of squares and $df = \text{rows}(Y) - \text{cols}(U)$ be the residual degrees of freedom, and T be the Cholesky decomposition such that $TT^T = (U^T U)^{-1}$. The relevant posterior distributions can be derived easily (see, for example, Gelman, Carlin, Stern and Rubin 1995, Chap. 7), and the following steps then provide draws from the posterior predictive distribution of missing Y values:

1. Generate a chi-square random deviate u with df degrees of freedom and define $\sigma_*^2 = SSE/u$.
2. Generate a vector $z = (z_1, z_2, \dots, z_p)$ of dimension $p = \text{rows}(B)$ of random normal deviates and define $\beta_* = B + \sigma_* Tz$.
3. Let U_{miss} denote the U -matrix for those with missing Y values. The imputed values are $Y_* = U_{\text{miss}} \beta_* + \sigma_* v$, where v is an independent vector of dimension $\text{rows}(U_{\text{miss}})$ of random normal deviates.

Binary Variable: When Y is a binary variable, fit a logistic regression model relating Y to U (most recently updated), $\text{logit}[\Pr(Y=1|U)] = U\beta$, using individuals with observed Y . The imputed values for Y are created through the following steps:

1. Let B denote the maximum likelihood estimates of β and V its asymptotic covariance matrix (negative inverse of the observed Fisher information matrix). Let T be the Cholesky decomposition of V (that is, $TT^T = V$). Generate a vector z of random normal deviates of dimension $\text{rows}(B)$. Define $\beta_* = B + Tz$.
2. Let U_{miss} denote the portion of U for which Y is missing. Define $P_* = [1 + \exp(-U_{\text{miss}} \beta_*)]^{-1}$. Generate a vector u , of dimension $\text{rows}(U_{\text{miss}})$ of uniform random numbers between 0 and 1. Impute 1 if a particular component of u is less than or equal to the corresponding component of P_* and impute 0 otherwise.

This approach results only in approximate draws from the posterior predictive distribution of the missing values as

the draws of the parameter β are from the asymptotic approximation of its actual posterior distribution. It is possible to draw from the actual distribution by modifying Step 1 using, for example, Sampling-Importance-Resampling (Rubin 1987b).

Mixed Variable: For Y , a mixed variable (that is, Y either takes the value zero or a continuous value), model the zero values by a 0-1 indicator to distinguish between 0 and non-zero values, and then model a normally distributed variable for the continuous portion of the distribution conditional on the indicator variable being equal to 1. That is, use a two stage approach: impute a one or zero using the logistic approach described above; and then restricting the sample to those with non-zero values, use the continuous variable approach described above to impute a continuous value to replace the just imputed value of 1.

Count Variable: For Y , a count variable, fit a Poisson regression model $Y \sim \text{Poisson}(\lambda)$ where $\log \lambda = U\beta$. The imputations for missing values in Y are created using the following steps:

1. Let B denote the maximum likelihood estimate of β , V its covariance matrix and T the Cholesky decomposition of V . Generate a vector z of random normal deviates of dimension $\text{rows}(B)$ and define $\beta_* = B + Tz$.
2. Let U_{miss} denote the portion of U for which Y is missing. Define $\lambda_* = \exp(U_{\text{miss}} \beta_*)$. Generate independent Poisson random variables with means as the elements of λ_* .

Polytomous Variable: For Y that can take k values, $j = 1, 2, \dots, k$, let $\pi_j = \Pr(Y=j|U)$. Fit a polytomous regression model relating Y to U where $\log = (\pi_j/\pi_k) = U\beta_j$ for $j = 1, 2, \dots, k-1$. Under the restriction $\sum_j \pi_j = 1$, it follows that $\pi_k = (1 + \sum_{j=1}^{k-1} \exp(U\beta_j))^{-1}$.

Let B denote the maximum likelihood estimate of the regression coefficients $(\beta_1', \beta_2', \dots, \beta_{k-1}')$, V be the asymptotic covariance matrix and T its Cholesky decomposition.

The following steps create imputations:

1. Define $\beta_* = B + Tz$ where z is a vector of random normal deviates of dimension $\text{rows}(B)$.
2. Let U_{miss} denote the rows of U with missing Y and let $P_i^* = \exp\{U_{\text{miss}} \beta_i\} / \{1 + \sum_i \exp(U_{\text{miss}} \beta_i)\}$ where β_i is the appropriate elements of β_* where $i = 1, 2, \dots, k-1$ and $P_k^* = 1 - \sum_i P_i^*$.
3. Let $R_0 = 0$, $R_j = \sum_i P_i^*$ and $R_k = 1$ be the cumulative sums of the probabilities. To impute values generate random uniform number u and take j as the imputed category if $R_{j-1} \leq u \leq R_j$.

Again, the imputation of mixed, count and categorical variables are from approximate posterior predictive distributions because the corresponding parameters are drawn from their asymptotic normal approximate posterior distributions.

REFERENCES

- BARNARD, J. (1995). Cross-Match Procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Unpublished Doctoral Thesis, University of Chicago, Department of Statistics.
- GELFAND, A.E., and SMITH, A.M.F. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.
- GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D.B. (1995). *Bayesian Data Analysis*. London. Chapman and Hall.
- GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-472.
- GELMAN, A., and SPEED T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of Royal Statistical Society*, B, 55, 185-188.
- GEMAN, S., and GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- HEERINGA, S.G., LITTLE, R.J.A. and RAGHUNATHAN, T.E. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan, Ann Arbor, Michigan.
- LI, K.H., MENG, X.L., RAGHUNATHAN, T.E. and RUBIN, D.B. (1991). Significance levels from repeated p values from multiply-imputed data. *Statistica Sinica*, 1, 65-92.
- LI, K.H., RAGHUNATHAN, T.E. and RUBIN, D.B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of American Statistical Association*, 86, 1065-1073.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.
- LITTLE, R.J.A., and SCHLUCHTER, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497-512.
- LIU, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis*, 53, 139-158.
- MADOW, W.G., NISSELSON, H., OLKIN, I. and RUBIN, D.B. (1983). *Incomplete Data in Sample Surveys*. 1,2, and 3, New York, Academic Press.
- MENG, X.L., and RUBIN, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103-111.
- OLKIN, I., and TATE, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32, 448-465.
- RAGHUNATHAN, T.E., and GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of American Statistical Association*, 90, 54-63.
- RAGHUNATHAN, T.E., and RUBIN, D.B. (1988). An application of Bayesian statistics using sampling/importance resampling to a deceptively simple problem in quality control. *Data Quality Control: Theory and Pragmatics*, (G.E. Liepins and V.R.R. Uppuluri, Eds). New York: Marcel Dekker.
- RAGHUNATHAN, T.E., and SISCOVICK, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.
- RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.
- RUBIN, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1978). Multiple imputation in sample surveys – A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.
- RUBIN, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- RUBIN, D.B. (1987b). The SIR-algorithm – A discussion of Tanner and Wong's. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91, 473-489.
- RUBIN, D.B., and SCHAFER, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceeding of the Statistical Computing Section of the American Statistical Association*, 83-88.
- SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*. New York: Chapman and Hall.
- SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of Royal Statistical Society*, B, 47, 1-52.
- SISCOVICK, D.S., RAGHUNATHAN, T.E., KING, I., WEINMANN, S., WICKLUND, K.G., ALBRIGHT, J., BOVBERG, V., ARBOGAST, P., KUSHI, L., COBB, L., COPASS, M.K., PSATY, B.M., RETZLAFF, B., CHILDS, M. and KNOPP, R.H. (1995). Dietary intake and cell-membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of American Medical Association*, 274, 1363-1367.

A Better Understanding of Weight Transformation Through a Measure of Change

JOHANE DUFOUR, FRANÇOIS GAGNON, YVES MORIN, MARTIN RENAUD and CARL-ERIK SÄRNDAL¹

ABSTRACT

The literature on longitudinal surveys of households offers several approaches for creating a set of final weights for use in data analysis. Most of these approaches depend on various procedures for modifying weights. Initial weights are often transformed into a set of intermediate weights in order to compensate for nonresponse, and then into a set of final weights, through poststratification, in order to adjust the sample. The literature includes a great deal of information about this approach but none of the studies has really looked closely at an approach for measuring the relative importance of these two steps in measuring the effectiveness of the numerous existing alternatives for creating intermediate weights. The objective of this paper is to study and measure the change (from the initial to the final weight) which results from the procedure used to modify weights. A breakdown of the final weights is proposed in order to evaluate the relative impact of the nonresponse adjustment, the correction for poststratification and the interaction between these two adjustments. This measure of change is used as a tool for comparing the effectiveness of the various methods for adjusting for nonresponse, in particular the methods relying on the formation of Response Homogeneity Groups. The measure of change is examined through a simulation study, which uses data from a Statistics Canada longitudinal survey, the Survey of Labour and Income Dynamics. The measure of change is also applied to data obtained from a second longitudinal survey, the National Longitudinal Survey of Children and Youth.

KEY WORDS: Nonresponse; Weighting; Calibration; Longitudinal survey; Measure of change.

1. INTRODUCTION

The literature contains many two-step approaches to transforming weights for household surveys. The first step involves an adjustment of the *initial weights* in order to compensate for nonresponse; the resulting weights are called *intermediate weights*. The second step produces the *final weights* through the process of poststratification, or more commonly through calibration (see Deville and Särndal 1992), in order to ensure that the final weights respect certain known population control totals. All of these weight modifications are designed to produce the "best possible set of final weights".

At Statistics Canada, longitudinal surveys of households also use this two-step approach in weighting, and the research work undertaken by the Agency leans in this direction. The U.S. Bureau of the Census "Survey of Income and Program Participation (SIPP)" (see Rizzo, Kalton and Brick 1996) also uses this type of approach.

Several methods are recommended in the literature for adjusting weights to compensate for nonresponse. Rizzo *et al.* (1996) compared the estimates obtained through several of these methods to estimates from independent sources. However, not many authors have done simulations or proposed tools for comparing the relative effectiveness of the methods in terms of their ability to reduce the nonresponse bias.

The main objective of this document is to study and measure the change (between initial and final weights) resulting from the adoption of a two-step procedure for modifying weights. Thus, a measure of change involving

four components is proposed in order to quantify the relative impact of the nonresponse adjustment, the correction for poststratification and the interaction between these two adjustments. The second objective is to use the measure of change to compare the effectiveness of the different nonresponse adjustment methods through a simulation study based on data from the Longitudinal Survey of Labour and Income Dynamics (SLID) and from the National Longitudinal Survey of Children and Youth (NLSCY). The longitudinal surveys are unique in that a great deal of information about respondents and nonrespondents to the latest wave is available from respondents to the previous waves. Thus, more complex methods can be used to adjust for nonresponse.

A general framework for the weighting of longitudinal surveys of households is presented in section 2. Then, the measure of change which will be used to quantify the stages of transformation between the initial and the final weights is presented in section 3. Section 4 addresses the nonresponse adjustment strategies contained in the literature. This is followed by sections 5 and 6, which contain the results of the studies based on the SLID and NLSCY. The last section presents the conclusions of this study.

2. GENERAL FRAMEWORK FOR LONGITUDINAL WEIGHTING

In a longitudinal survey of households, individuals in the initial sample are followed over time, and are referred to as *longitudinal individuals*. This set of individuals is the one

¹ Johane Dufour, François Gagnon, Yves Morin, Martin Renaud and Carl-Erik Särndal, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

which will be used in the studies presented in this document. They are referred to as the “reference unit”. This section provides an overview of the steps followed in order to modify the initial weight for longitudinal individuals into a final weight.

2.1 Initial Weights

$U = \{1, \dots, k, \dots, N\}$ is a finite population. We are interested in variable y (the variable of interest), whose value for the k -th unit is recorded as y_k . The objective is to estimate the total $Y = \sum_U y_k$. Let w_{0k} be the initial weight for all $k \in s$ units, where s is the longitudinal sample. In the absence of nonresponse, the set of initial weights $\{w_{0k}: k \in s\}$ yields the $\hat{Y} = \sum_s w_{0k} y_k$ estimator for Y . In this case we assume that the w_{0k} are normalized in order to ensure that $\sum_s w_{0k} = N$. Although \hat{Y} is unbiased for Y , \hat{Y} has the drawback of not incorporating any ancillary information in the form of known control totals for poststrata.

2.2 Nonresponse Adjustment and Intermediate Weights

Most surveys have to deal with nonresponse. Two approaches are often used to compensate for this: imputation and the correction of the initial weights of respondents through an adjustment factor. The latter is the one more commonly used in household surveys to compensate for total nonresponse, while imputation is often preferred when dealing with partial nonresponse. Total nonresponse reduces the size of the sample since the y_k value is only available for $k \in r$, where $r \subset s$ is the set of the m responding units. For this reduced set of data, the initial w_{0k} weights are, on average, too small and we have $\sum_r w_{0k} < N$. The estimator $\hat{Y}' = \sum_r w_{0k} y_k$ is not admissible since it systematically underestimates Y .

Weight adjustment is often chosen in order to compensate for total nonresponse in household surveys. A common method of adjusting weights involves constructing Response Homogeneity Groups (RHGs). These are designed so that each one is comprised of reference units having a similar probability of response. Then, within each RHG, an adjustment factor equal to the inverse of the RHG's response rate (weighted or not) is calculated. For each respondent unit k , the adjustment for nonresponse involves multiplying w_{0k} by the RHG's adjustment factor. This operation results in a set of intermediate weights $\{w_{1k}: k \in r\}$, where $\sum_r w_{1k} = N$. With these weights, we can construct the estimator $\hat{Y}'' = \sum_r w_{1k} y_k$, which eliminates the underestimation which is characteristic of $\hat{Y}' = \sum_r w_{0k} y_k$. As in the case of the set of initial weights, the main drawback with this set is that it fails to incorporate the ancillary information available for poststrata.

2.3 Poststratification and Final Weights

A widely-used practice in household surveys involves modifying the intermediate weights through poststratification, or, more commonly, through calibration, so that

the sum of the final weights on the set of respondents will correspond to the known population counts. Thus, poststratification produces a set of final weights $\{w_{2k}: k \in r\}$, which incorporates the ancillary information and which is also consistent with the control totals for the poststrata. In this case, the final weights in each poststratum p confirm $\sum_{r_p} w_{2k} = N_p$, where N_p is the known element and r_p is the set of respondent units in the p -th poststratum. It follows that $\sum_r w_{2k} = N$. Demographic and geographic variables are frequently used to define poststrata. The choice of poststrata, which must be sufficiently large, is limited by the availability of control totals. Several methods may be used to calibrate the intermediate weights to the selected control totals.

3. MEASURE OF CHANGE FROM INITIAL TO FINAL WEIGHTS

In this section, a measure of the change between initial and final weights is presented so to better understand the effect of the weight modification procedure. The breakdown of this measure into four components makes it possible to quantify the effect of each of the weighting steps described in section 2. These components will be used in sections 5 and 6 in the comparison of various methods for adjusting weights for nonresponse.

If the initial weights are normalized so that $\sum_s w_{0k} = N$, and if $r \subset s$, then the three sets of weights described in section 2 confirm the following relations:

$$\sum_r w_{0k} < N, \sum_r w_{1k} = N, \sum_r w_{2k} = N.$$

Let

$$\bar{w}_{01} = \frac{\sum_r w_{1k}}{\sum_r w_{0k}} \text{ and } \bar{w}_{02} = \frac{\sum_r w_{2k}}{\sum_r w_{0k}}.$$

The ratio \bar{w}_{01} measures the average change in the intermediate weight set in relation to the initial weight set. As total nonresponse becomes more pronounced, \bar{w}_{01} shifts farther away from the value of 1, which is only obtained in the absence of nonresponse. The ratio \bar{w}_{02} represents the average change in the set of final weights in relation to the set of initial weights.

The \bar{w}_{01} and \bar{w}_{02} ratios measure the average change in weight. To measure an individual change in weight, we define, for every $k \in r$, $r_{01k} = w_{1k} / (w_{0k} \bar{w}_{01})$, and $r_{02k} = w_{2k} / (w_{0k} \bar{w}_{02})$. These quantities vary around 1. More specifically, their weighted averages equal 1:

$$\frac{\sum_r w_{0k} r_{01k}}{\sum_r w_{0k}} = \frac{\sum_r w_{0k} r_{02k}}{\sum_r w_{0k}} = 1.$$

The r_{01k} and r_{02k} quantities will be useful for measuring individual weight changes.

The total weight change, from the set of initial to final weights, going through the set of intermediate weights, can be calculated by a measure of change, also called *distance*. Here, D is the following measure of change:

$$D = \frac{\sum_r w_{0k} \left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2}{\sum_r w_{0k}}.$$

In fact, D is a weighted average of the following individual weight change factors:

$$\left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2 = \left(\frac{w_{2k}}{w_{1k}} \frac{w_{1k}}{w_{0k}} - 1 \right)^2.$$

The measure of change D breaks down into four components, as set out in the following equation:

$$D = R_{01} + R_{12} + R_{\text{int}} + G$$

where:

$$R_{01} = \bar{w}_{02}^2 \frac{\sum_r w_{0k} (r_{01k} - 1)^2}{\sum_r w_{0k}},$$

$$R_{12} = \bar{w}_{02}^2 \frac{\sum_r w_{0k} (r_{02k} - r_{01k})^2}{\sum_r w_{0k}},$$

$$R_{\text{int}} = 2\bar{w}_{02}^2 \frac{\sum_r w_{0k} (r_{01k} - 1)(r_{02k} - r_{01k})}{\sum_r w_{0k}} \text{ and}$$

$$G = (\bar{w}_{02} - 1)^2.$$

It should be noted that the measure of change D is always positive, equality being at zero when the two following conditions are met:

- (i) absence of nonresponse ($r = s$ and $w_{1k} = w_{0k}$ for all k),
- (ii) absence of poststratification effect on the intermediate weights ($w_{2k} = w_{1k}$ for all k).

A high nonresponse rate would tend to increase the value of the measure of change D since in such a case, w_{1k} is generally much larger than w_{0k} .

R_{01} measures the individual weight changes which result from going from the initial to the intermediate set. Later, we will see that the component R_{01} is somehow associated with the quality of the nonresponse model and that a large R_{01} value is preferable. R_{12} measures the individual weight changes which result from going from the intermediate to

the final set. R_{int} measures the interaction between the two types of change and G measures the change in average weight between the initial and final sets.

In addition to its interpretation as a distance, the measure of change D can also be interpreted as a mean square error of changes w_{2k}/w_{0k} in relation to 1, and in relation to the distribution defined by all the w_{0k} . From this perspective, the component G corresponds to the bias squared (or the square of the difference between the \bar{w}_{02} average of w_{2k}/w_{0k} and 1), while the sum of the other three components corresponds to the variance. In the simplest case, where a nonresponse adjustment is calculated using a single RHG, and where no poststratification is applied, we have $w_{0k} = N/n$ for all $k \in s$ (in the case of a size n simple random selection) and $w_{1k} = w_{2k} = N/m$ for all $k \in r$, (where the nonresponse adjustment factor is n/m , i.e., the inverse of the response rate). We then have $D = G = \{(n/m) - 1\}^2$ and $R_{01} = R_{12} = R_{\text{int}} = 0$.

Some significant conclusions may be drawn from looking at the relative importance of R_{01} , R_{12} and R_{int} . If R_{01} is high at the same time that R_{12} is not very high, the survey is one in which the nonresponse adjustment creates significant individual changes in weights, while poststratification only results in a slight change in individual weights. However, when R_{12} is high, poststratification brings about very large individual changes. The results presented in sections 5 and 6 will show that R_{01} can be used to compare the effectiveness of various nonresponse adjustment methods. As well, the sign of R_{int} indicates whether the two types of individual change are moving in the same direction ($R_{\text{int}} > 0$) or in opposite directions ($R_{\text{int}} < 0$). In reality, we expect R_{int} to be very small, if not negligible.

4. NONRESPONSE ADJUSTMENT STRATEGIES

The literature contains several methods for adjusting weights (including the method described in section 2.2) to compensate for nonresponse. Another method, which is frequently used in longitudinal surveys, involves adjusting weights in accordance with the inverse of the predicted probability of response obtained through a logistic regression. We also find methods of adjustment based on calibration, which use marginal distributions of the initial sample or of the population. Singh, Wu and Boyer (1995) used this approach in order to derive a method of adjustment capable of producing coherent estimates in longitudinal surveys from one wave to the next. Deville (1998) recommended a method of correction for nonresponse by calibration or balanced sampling. For a review of nonresponse adjustment methods, refer to Kalton and Kasprzyk (1986), Platek, Singh and Tremblay (1978), Chapman, Bailey and Kasprzyk (1986) and to Little (1986). In this document, only methods relying on the creation of RHGs are considered.

4.1 Formation of RHGs

In most surveys, aside from a few stratification variables from the sample frame, very little information is available about non-respondents. Therefore, the choice of RHGs is very limited and the strata are often used as RHGs. In these cases, the assumption is that the probability of response is the same for all units in a given stratum. However, in longitudinal surveys, a great deal of information about respondents and non-respondents in the current wave is available from the responses provided in the previous waves. This information can then be used to create RHGs within which the assumption of a uniform response mechanism is plausible. This leads to a better nonresponse adjustment and, therefore, a reduction in the risk of introducing a nonresponse bias into the estimates.

4.1.1 Method for the Selection of Variables for the Formation of RHGs

By definition, an RHG is formed from a set of variables capable of predicting the propensity to respond. If the set of variables which is defined at the outset is too large, univariate tests may be used to isolate the most important variables to distinguish the characteristics of respondents from those of nonrespondents. With this set of important variables, a selection method may then be applied for retaining the best variables for explaining the propensity to respond. Two of the current variable selection methods are: the Logistic Regression Model (LR) and the Segmentation Model (SM).

4.1.1.1 Logistic Regression

Under the LR method, the combined use of the “fact of having responded to the survey or not” as a dependent variable, standardized weights and the “stepwise” procedure result in a list of the most significant dichotomic variables for explaining the propensity to respond. As a general rule, RHGs are created according to 2^q possible combinations, based on a set of q explanatory variables used. The LR is often referred to as the symmetrical approach. However, if certain additional constraints are applied when the RHGs are created, this could reduce their numbers. For instance, we could require a minimum number of reference units (n) and a response rate (RR) (weighted or not weighted) greater than a certain level in each of the RHGs. Kalton and Kasprzyk (1986) encourage the use of such constraints in order to avoid increasing the variance associated with extreme weights. However, these constraints may reduce the effectiveness of the nonresponse adjustment and result in an increase in the bias. When an RHG does not meet one of these constraints, it has to be combined with another RHG. The combination of RHGs continues until all of the RHGs meet the additional constraints imposed. This leads to $2^q - J$ valid combinations, where J represents the reduction resulting from the combination of RHGs.

For instance, in Figure 1, $2^q = 8$ RHGs are created on the basis of $q=3$ explanatory variables. The shaded boxes in Figure 1 represent the RHGs. An adjustment factor is calculated within each RHG and the weight w_{ok} of each reference unit is then adjusted, accordingly.

4.1.1.2 Segmentation Model

The SM method, which is referred to as non-symmetrical, is based on the CHAID (Chi-square Automatic Interaction Detection) algorithm developed by Kass (1980). It divides the sample into sub-groups according to the response rate of the explanatory variables by using a Chi-square test. The segmentation process continues until a significant explanatory variable is found. The final sub-groups created through the SM become the RHGs, for which the nonresponse adjustments are calculated. As in the case of the LR, additional constraints may be imposed.

In Figure 1 we see that the SM method divided the sample into several RHGs based on the different explanatory variables. The RHGs are once again represented by the shaded boxes. The segmentation continues until it is no longer possible to find explanatory variables.

4.1.2 Nonresponse Adjustment Factor

Whether the RHGs are formed by relying on the LR or the SM, a uniform response mechanism is assumed within each RHG. Thus, the nonresponse adjustment factor is given by the inverse of the response rate (weighted by w_{ok} or not weighted) for the RHG.

5. EMPIRICAL STUDY BASED ON THE SURVEY OF LABOUR AND INCOME DYNAMICS (SLID)

Data from the SLID were used for an empirical study designed to compare the effectiveness of the LR and SM. The SLID is a longitudinal survey of households that started in 1993; one of its objectives is to provide information on the economic well-being of Canadian society (see Lavigne and Michaud 1998).

These two methods were tested through a simulation by analyzing some variables of interest and various domains. The components of the measure of change, the absolute and relative biases and the variances were studied.

5.1 Description of the Empirical Study

The first step in the empirical study was to estimate the probability of response to the first wave of the survey for each of the units in the longitudinal sample. Variables which could potentially explain the propensity to respond (based on a preliminary interview) were used to form a very large number of RHGs. All of the individuals in the sample were assigned to an RHG on the basis of the values of the explanatory variables. A probability of response was then estimated for each RHG on the basis of the weighted response rate. Then, only the respondents and their

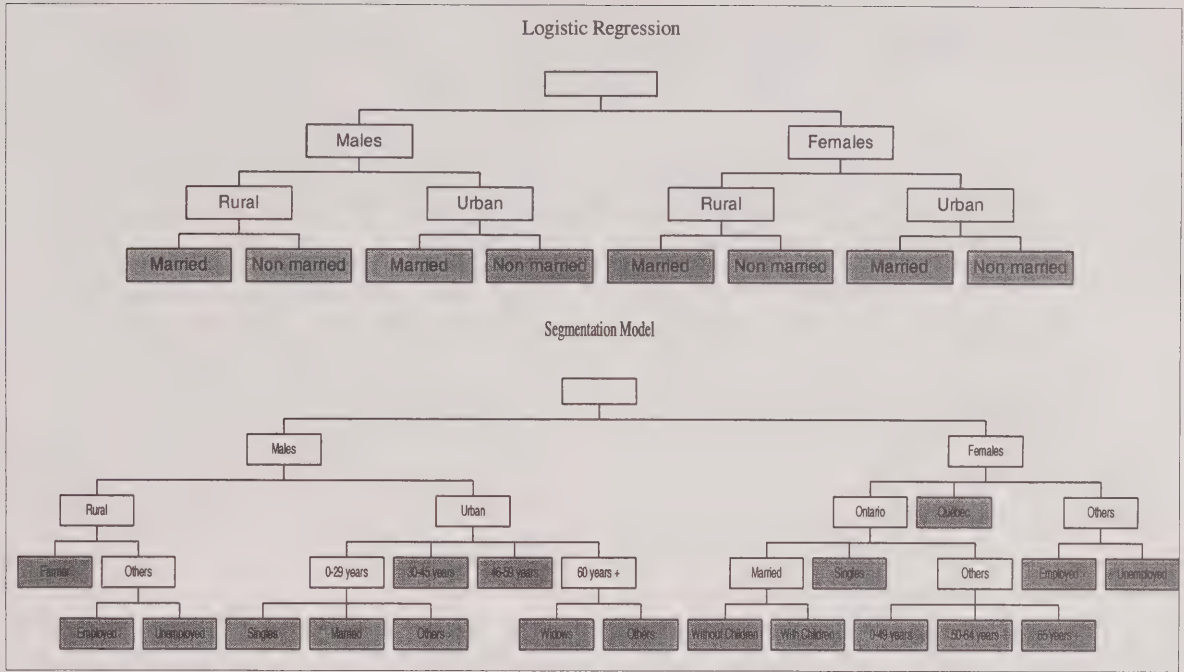


Figure 1. Depiction of the Formation of RHGs by Method

probability of response were retained in the reference sample for the simulation. Nonresponse was then generated for the reference sample through Poisson sampling. This procedure, illustrated in Figure 2, was independently repeated 100 times, thus creating 100 sets of respondents and non-respondents. The average response rate for each repetition was around 90%, which was the rate observed in the first wave of the SLID.

For each of the 100 repetitions, a nonresponse adjustment was done using the LR method to create the RHGs. Similarly, a nonresponse adjustment was done using the SM to create RHGs for each of the first 20 repetitions. With the SM approach, the number of repetitions was limited to 20, given the stability of the results and since several manual interventions and the use of a specific software package (in our case: Knowledge Seeker – ANGOSS Software 1995) were required.

Several variants of the variable selection method were studied:

a) LR_{*i*}, where *i* represents, out of the 100 repetitions, the approximate average of the number of RHGs generated through the LR method. In this study, *i*=4, 16, 40, 60. For instance, for LR₄₀, the *q*=6 most important explanatory variables for the propensity to respond were first identified. The RHGs were then formed using the ($2^q - J$) valid combinations of these *q*=6 explanatory variables. The imposition of additional constraints (*n* > 30 and RR > 50%) in each RHG led to the re-grouping of some RHGs. On average, out of 100

repetitions, 24 RHGs had to be regrouped (*J*=24) and a total $2^q - J = 2^6 - 24 = 40$ RHGs were formed, hence the LR₄₀ designation. In the simulation study, LR_{*i*}, where *i*=4, 16, 40, 60 RHGs corresponds, respectively to *q*=2, 4, 6, 8 explanatory variables.

- b) SM_{*i*}, where *i* indicates the approximate average in the first 20 repetitions of the number of RHGs generated through the SM method. In this study, *i*=16, 25, 40. For example, for SM₁₆, one SM was used with a significance level *p* of 0.0001. After the imposition of the same additional constraints as for the LR, an average 16 RHGs were created. SM_{*i*}, where *i*=16, 25, 40 RHGs corresponds, respectively, to the significance levels of 0.0001; 0.0005; 0.0025. The higher the level used, the easier it is to identify the significant differences, which makes it possible to achieve a more detailed segmentation and, hence, a greater number of RHGs.
- c) A method with a single RHG (1_RHG) was also used for comparison purposes. This method involves defining the entire sample as a single RHG for each of the 100 repetitions. It should be noted that this method is only effective if the response mechanism is uniform within the entire sample, which is rarely the case.

At first, the initial weights were normalized so that $\sum_k w_{0k} = N$, in order to eliminate the effect of under-coverage and to better isolate the effect of nonresponse. Thus, *G* will only measure the average change caused by the nonresponse adjustment.

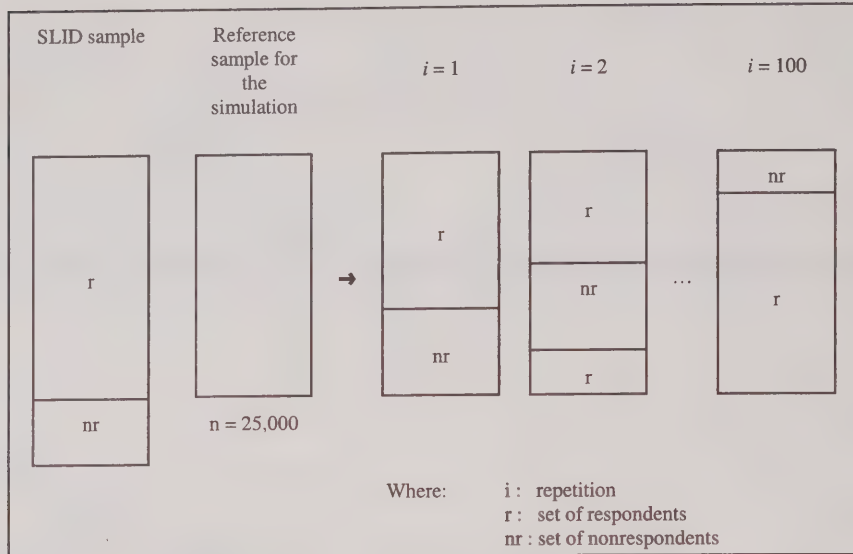


Figure 2. Illustration of the Simulation Process

Once the initial weights were normalized, each set of final weights was then the result of a two step process: a nonresponse adjustment (based on one of the eight methods mentioned: 1_GRH, LR_i, where $i = 4, 16, 40, 60$ and SM_i, where $i = 16, 25, 40$) and a same poststratification (14 age-sex groups by province).

5.2 Analysis of the Results of the Empirical Study

For each of the methods discussed in the previous section, the components of the measure of change D were studied. Also, the average, absolute and relative nonresponse bias and the average variance of the estimates were analyzed.

5.2.1 Measure of Change (D)

Table 1 presents the average value of D and its components for each of the M repetitions (where $M=100$ for the LR and $M=20$ for the SM) as well as the percentage contribution of each element to the average value of D . We observe, in the first place, that for the 1_GRH method, R_{01}

is nil since one single nonresponse adjustment was made to the set of respondents. Thus, $w_{1k} = \alpha w_{0k}$, where α is a constant, so $r_{01k} = 1$ for every $k \in r$ and $R_{01} = 0$. We also observe that D increases as the number of RHGs increases, irrespective of whether the LR or SM method is used. Thus, the more RHGs there are to compensate for nonresponse, the greater the total change to which the weights are subjected. In addition, the values of D are higher for the SM than for the LR.

For the LR and the SM, the contribution of R_{01} to the measure of change increases as the number of RHGs increases, since nonresponse is more readily targeted as the number of RHGs increases. Consequently, the nonresponse adjustment often becomes more important and, thereby, the weights vary more and more. In addition, the contribution of R_{01} to the measure of change is much more important with the SM than with the LR. This indicates that the SM seems to be better at modeling nonresponse and isolating the specific trends of the LR.

Table 1
Average Value of D on Repetitions, for each Component and their Contribution (as a %) to the Measure of Change for each of the Eight Nonresponse Adjustment Methods

Method	D	R_{01} ($\times 10^{-3}$)	R_{01}/D (%)	R_{12} ($\times 10^{-3}$)	R_{12}/D (%)	R_{nt} ($\times 10^{-5}$)	R_{nt}/D (%)	G ($\times 10^{-2}$)	G/D (%)
1_RHG	0.012135	0.00	0.00	1.17	9.66	0.00	0.00	1.11	90.34
LR_4	0.012952	0.78	6.04	1.10	8.49	0.06	0.01	1.11	85.46
LR_16	0.013809	1.66	11.97	1.00	7.31	3.76	0.54	1.11	80.19
LR_40	0.014426	2.32	16.02	0.96	6.66	4.02	0.55	1.11	76.77
LR_60	0.014948	2.85	19.00	0.95	6.35	3.75	0.49	1.11	74.15
SM_16	0.015712	3.42	21.33	0.97	6.19	3.40	0.43	1.11	72.05
SM_25	0.016713	4.44	26.02	0.95	5.73	2.95	0.36	1.11	67.89
SM_40	0.018202	5.97	32.37	0.95	5.23	1.20	0.14	1.11	62.26

As for R_{12} , it is almost constant, regardless of which method and number of RHGs are used. However, despite the fact that it changes very little, its contribution to the measure of change diminishes as the number of RHGs increases. This is due to the fact that there is more variation in the weights with a nonresponse adjustment, and the modifications which poststratification creates in the weights are less and less important as the number of RHGs increases.

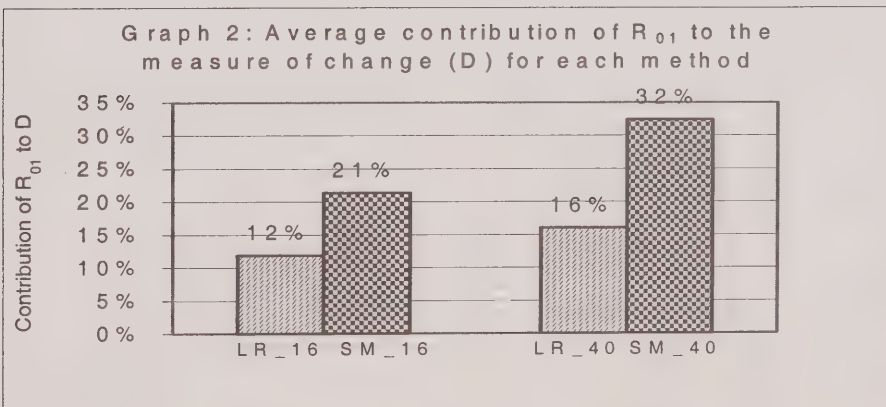
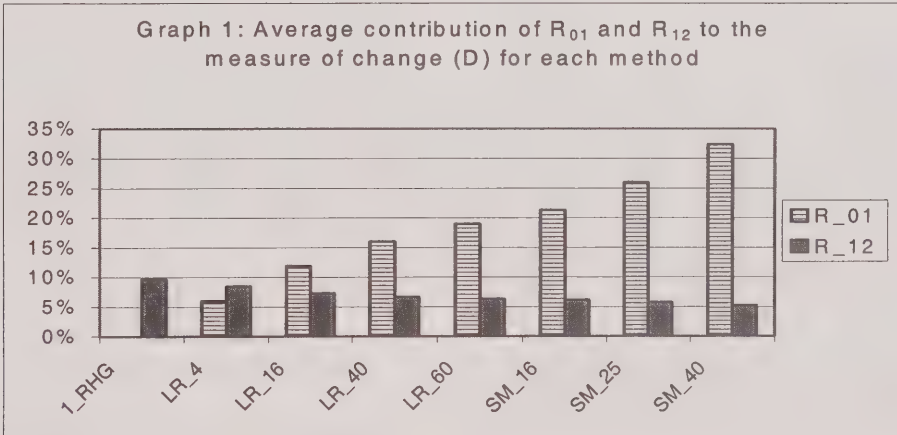
In the case of R_{int} , its value is negligible and its contribution to the measure of change is very small. This means that the interaction between the nonresponse adjustment and poststratification is practically nil.

Finally, G remains constant, irrespective of which method and how many RHGs are used. As with R_{12} , the contribution of G to the measure of change diminishes as the number of RHGs increases. A larger number of RHGs is better at targeting nonresponse, thereby causing more variations in the set of intermediate weights.

Since, with all of these methods, G is constant, R_{int} is close to zero and R_{12} is nearly constant, it is clear that the variations in D are mostly influenced by the variations in R_{01} .

Graph 1 shows the average contribution in percentage of R_{01} and R_{12} to the measure of change. For LR and SM, the contribution of R_{01} increases with the number of RHGs while that of R_{12} diminishes. Also, the contribution of R_{01} is greater for SM than for LR, while that of R_{12} is less for SM than for LR. The profile of the contribution of R_{01} is the same as the profile of D (Table 1). This confirms that the variations in the measure of change are mainly due to the variations in R_{01} .

Graph 2 shows the comparison between the LR and SM in terms of the average percentage contribution in percentage of R_{01} to D . For a given number of RHGs, R_{01} contributes to a larger percentage of D through the SM method than through the LR method. This means that individual changes in the weights between the initial and intermediate sets are greater for SM than for LR.



5.2.2 Relative and Absolute Biases

The Relative Bias (RB) and the Absolute Bias (AB) were used to compare the performance of LR relative to SM in reducing the nonresponse bias:

$$RB_i = 100 \left(\frac{\hat{Y}_i - Y}{Y} \right) \text{ and } AB_i = \hat{Y}_i - Y;$$

where \hat{Y}_i is the estimate of the variable of interest obtained for the i -th repetition, $i = 1, 2, \dots, M$, $M=100$ for the LR, $M=20$ for the SM and Y is the total for the variable of interest obtained from the reference sample.

The Average Relative Bias (ARB) and the Average Absolute Bias (AAB) are calculated by taking, respectively, the average of the RB and the AB for all repetitions:

$$ARB = \frac{1}{M} \sum_{i=1}^M RB_i \text{ and } AAB = \frac{1}{M} \sum_{i=1}^M AB_i$$

where $M=100$ in the case of the LR and $M=20$ in the case of the SM.

For the 100 repetitions, national estimates were produced for the following three variables: "person living, or not, in a family whose revenue is less than the Low Income Cutoff (LICO)", "Individual Total Income (TI)" and "Individual Wages and Salaries (WS)". The ARB for each estimate was calculated for the eight methods under study. Given the large sample size, the low nonresponse rate (10%) and the fact that a large number of control totals was used for poststratification, the ARB is very small (see Table 2) for each of the methods used.

In Table 2 we see that, for each of the three variables, the ARB is more or less constant for the SM, irrespective of how many RHGs are used. Also, for the LR, the ARB for the TI and SW is more or less constant not withstanding the number RHGs used. On the other hand, for the LICO, the ARB for method LR_4 is much smaller than the ARB for the other three LR methods. This could be due to the fact that the LICO is a variable derived from several other variables, unlike the TI and the SW, which are observed variables. The ARB for the three variables for method 1_RHG is much larger than the ARB produced by the SM and the LR, except for the LICO, since in this case the ARB is more or less equivalent to the ARB of the LR. Thus, it appears that method 1_RHG does not perform as well as the SM and the LR. In the best case, it is more or less equivalent to LR. Unlike SM, we observe that the progression

of ARB is not strictly downwards for the LR, as the number of RHGs increases.

Despite the fact that the ARB is minimal for the variables studied for Canada, it can increase rapidly for small domains. In this study, other domains were also reviewed. Although some variances were observed in several of these cases, it seems that the ARB for the SM is generally smaller than the ARB for the LR and the method 1_RHG. A more detailed study of a larger number of interest and domain variables would be beneficial for corroborating these conclusions.

As previously indicated, the individual changes in the weights caused by the nonresponse adjustments are greater for the SM than for the LR (see Graph 2). This would suggest that the SM is more effective in reducing the nonresponse AB for a fixed number of RHGs. Graph 3 confirms this observation, showing that the AAB for the LICO is smaller through the SM than through the LR method.

5.2.3 Variance Estimates

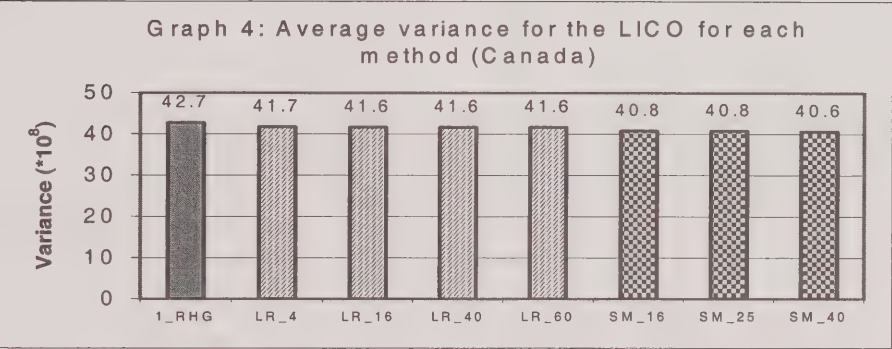
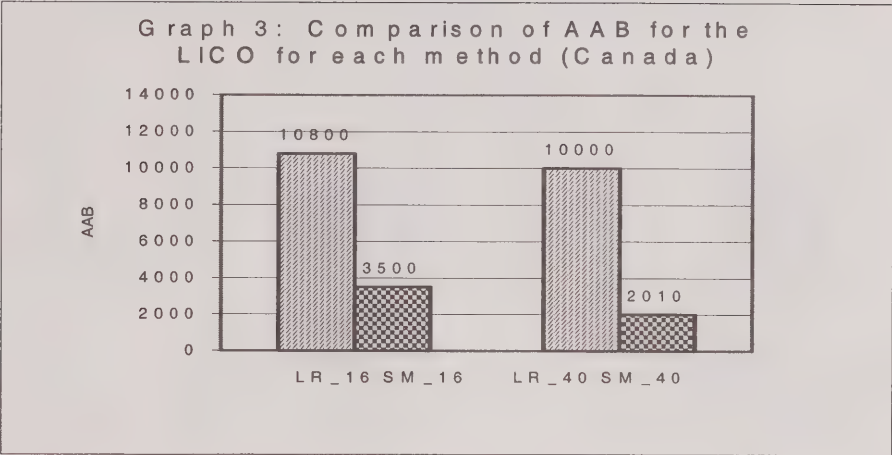
Variance estimates were produced for the three variables of interest through the Jackknife method. For LICO (Graph 4), the average variance of estimates is approximately the same, regardless of the method used. However, there is a slight decrease when the number of RHGs increases, for both the LR and the SM. Also, based on the empirical study, average variance estimates for the SM are slightly smaller than for the LR. Therefore, the larger dispersion in the weight (a higher value for D) does not entail an increase in variance.

6. APPLICATION TO THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH (NLSCY) DATA

In this section, most of the analyses done with the help of the LR and SM in the empirical study with data from the SLID are reproduced with the information obtained from the NLSCY. Just like the SLID, the NLSCY is a longitudinal survey of households. It started in 1994 and is designed to collect information for analyzing policies and developing programs addressing critical factors affecting the development of children in Canada (see Michaud, Morin, Clermont and Laflamme 1998).

Table 2
ARB (as a %) for Different Variables Based on the Methods – Canada

Variable	STUDIED METHOD							
	1_RHG	LR_4	LR_16	LR_40	LR_60	SM_16	SM_25	SM_40
LICO	0.37	0.15	0.43	0.37	0.31	0.14	0.12	0.08
TI	-0.32	-0.09	-0.06	-0.05	-0.06	-0.006	-0.005	0.002
WS	-0.44	-0.13	-0.15	-0.19	-0.14	-0.10	-0.09	-0.09



6.1 Description and Analysis of the Results of the Application

The following methods were used for this study: LR_{*i*}, where *i*=4, 14, 41, 70 with, respectively *q*=2, 4, 6, 8 variables, and SM_{*i*}, where *i*=19, 36 with significance levels of 0.001 and 0.005, respectively. The same two constraints imposed for the SLID were re-applied when the RHGs were created. The same poststratification was used (22 age-sex groups by province) for each of the methods under study.

Unlike the empirical study based on the SLID, only the data collected in the first two waves of the NLSCY were used. There was no simulation and the initial weights were not normalized ($\sum_s w_{0k} = \hat{N} < N$). It should be noted that the undercoverage of the NLSCY is around 13% and its nonresponse is around 8%.

The conclusions drawn from the results presented in Table 3 are similar to those obtained in the simulation

(Table 1). However, we observe that the relative contribution by R_{01} to the measure of change is weaker for the NLSCY than for the SLID. This result indicates that the nonresponse adjustment of the SLID produces larger individual changes in the weights, thereby resulting in a larger contribution by R_{01} . Therefore, the nonresponse adjustment in the case of the NLSCY had no significant effect on the individual changes in the weights, contrary to what was observed in the case of the SLID.

The relative contribution by R_{12} to the measure of change is higher for the NLSCY than for the SLID. This result indicates that the more refined poststratification of the NLSCY results in greater individual changes in the weights, which translates into a greater contribution of R_{12} . Therefore, the NLSCY benefits a great deal from poststratification, which is less important for the SLID.

Table 3

Value of D , for each Component, and of their Contribution (as a %) to the Measure of Change for each of the Six Nonresponse Adjustment Methods

Method	D	R_{01}	R_{01}/D (%)	R_{12}	R_{12}/D (%)	R_{int} ($\times 10^{-4}$)	R_{int}/D (%)	G	G/D (%)
LR_4	0.1475	0.0052	3.51	0.0369	25.05	-4.63	-0.31	0.1058	71.76
LR_14	0.1497	0.0075	5.00	0.0367	24.69	-5.50	-0.37	0.1058	70.68
LR_41	0.1530	0.0112	7.29	0.0369	24.13	-9.16	-0.60	0.1058	69.18
LR_70	0.1564	0.0144	9.21	0.0362	23.13	-0.19	-0.01	0.1058	67.67
SM_19	0.1608	0.0187	11.63	0.0371	23.07	-8.24	-0.51	0.1058	65.81
SM_36	0.1640	0.0220	13.41	0.0373	22.76	-11.30	-0.69	0.1058	64.52

With respect to R_{int} , as with the SLID, its contribution to the measure of change is negligible. Contrary to the SLID, the sign of R_{int} is negative, which means that the interaction between R_{01} and R_{12} is negative.

With respect to G , as in the case of the SLID, it is the key source of contribution to the measure of change. In the case of the NLSCY, G not only includes the average change in weight resulting from the nonresponse adjustment, but also the average change in weight resulting from the correction for undercoverage through poststratification.

When all of these results are compared, it becomes evident that the two surveys are very similar since $R_{int} \approx 0$ and the sum of the contributions to the measure of change of R_{01} and R_{12} is around 35% in both cases. However, the NLSCY is also very different from the SLID since R_{12} predominates in the former one, while R_{01} predominates in the latter.

Just as with the SLID, D increases with the number of RHGs and this measure is greater for the SM than for the LR. In fact, the value of D is greater for the NLSCY than for the SLID, mainly because of the NLSCY under-

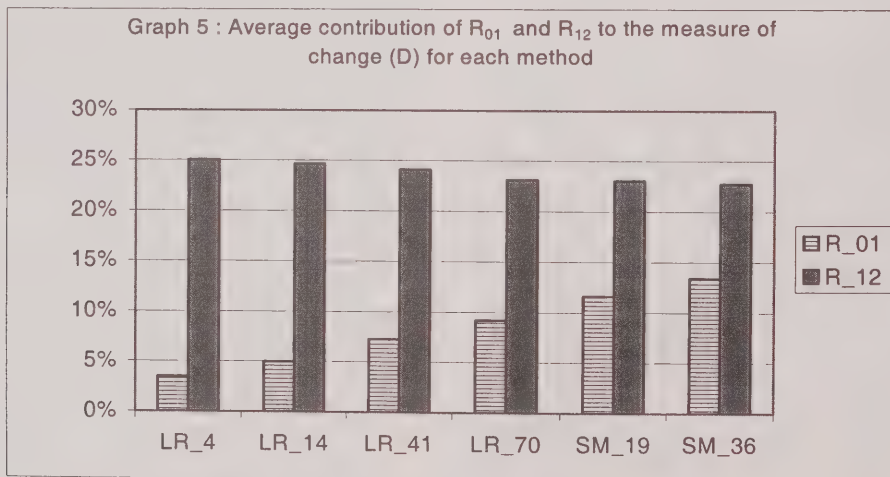
coverage, which results in an increase in G and, therefore, in D .

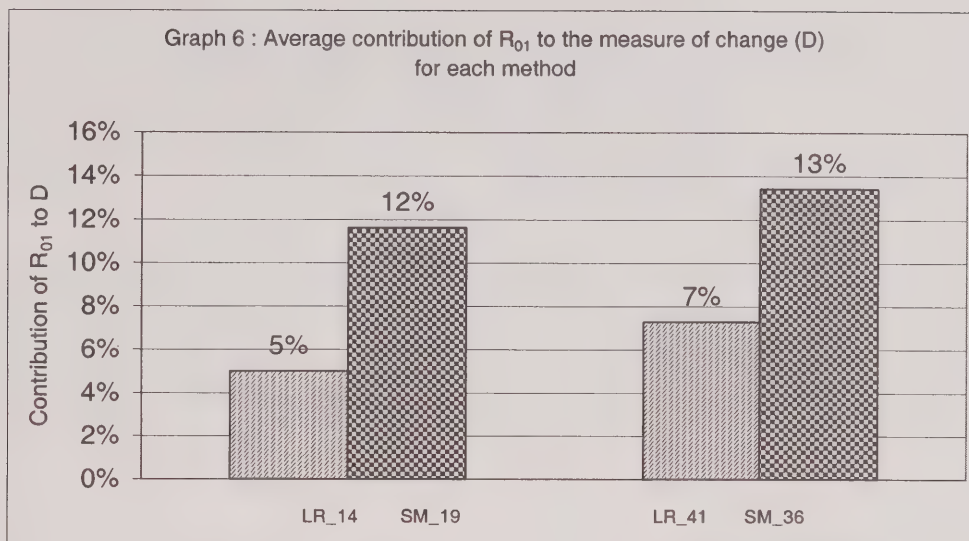
The average contribution of R_{01} for the LR and the SM increases with the number of RHGs, whereas that of R_{12} diminishes (Graph 5). The contribution of R_{01} is also greater for the SM than for the LR, unlike the contribution of R_{12} , which is smaller for the SM than for the LR.

As was observed with the empirical study, the profile of the contribution of R_{01} to the measure of change is the same as that of the measure itself. This shows that the variations in D depend directly on R_{01} .

Graph 6 enables us to compare the LR and the SM, presenting the average contribution of R_{01} to the measure of change for the methods with an essentially equivalent number of RHGs. As with the SLID, the results indicate that nonresponse seems to be better targeted with the SM than with the LR method.

Unlike the SLID simulation study, the bias was not evaluated since no external source of data was available for evaluation purposes.





7. CONCLUSION

This document highlights the fact that the choice of RHGs and method for defining them depends on the: i) availability of ancillary information, ii) need to reduce the nonresponse bias for all estimates, and iii) time and operational constraints. The empirical study, as well as the NLSCY data, showed that the SM method appears to be better than the LR one in reducing the nonresponse bias. The results also demonstrated that the proposed measure of change can be a very useful tool for comparing different weighting strategies.

In particular, it would appear that, as the value of R_{01} increases, the reduction of the bias obtained from using RHGs increases. Given the difficulty in obtaining a reliable estimate of the nonresponse bias in a survey, the relationship identified between the size of R_{01} and the decrease in the bias suggests that R_{01} should be used as a tool for evaluating nonresponse adjustment methods. This requires that R_{01} first be determined for different RHG sets. Then, the set with the highest R_{01} value is likely to be more effective than the other alternatives in reducing the nonresponse bias for most of the variables of interest.

The measure of change presented could also be used to compare the different calibration strategies. In this case, the nonresponse adjustment could remain the same for all of the poststratification methods under study. A detailed study of the behaviour of R_{12} could be done and would no doubt lead to certain conclusions, as this study did about R_{01} . This type of study would not necessarily have to be restricted to the longitudinal context but could quite readily be done with a cross-sectional study. Also, the measure of change could be useful in evaluating different nonresponse adjustment methods in cross-sectional surveys.

ACKNOWLEDGMENTS

The authors would like to thank M. Hladky, M. Latouche, C. Nadeau and N. Tremblay for their important contributions to this project.

REFERENCES

- ANGOSS SOFTWARE (1995). Knowledge Seeker IV for Windows – User's Guide. ANG OSS Software International Limited.
- CHAPMAN, D.W., BAILEY, L. and KASPRZYK, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology*, 12, 161-179.
- DEVILLE, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *1998 Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 103-110.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.
- LAVIGNE, M., and MICHAUD, S. (1998). General aspects of the Survey Labour and Income Dynamics. SLID research paper, Statistics Canada, catalogue number 98-05.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 137-139.

- MICHAUD, S., MORIN, Y., CLERMONT, Y. and LAFLAMME, G. (1998). Issues in the design of a survey to measure child development: The Experience of the Canadian National Longitudinal Survey of Children and Youth. Statistics Canada, Internal document.
- PLATEK, R., SINGH, M.P. and TREMBLAY, V. (1978). Adjustment for nonresponse in surveys. *Survey Sampling and Measurement*. N.K. Namboodiri, Ed. Academic Press, 157-174.
- RIZZO, L., KALTON, G. and BRICK, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.
- SINGH, A. C., WU, S. and BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 396-401.

Sampling and Weighting a Survey of Homeless Persons: A French Example

PASCAL ARDILLY and DAVID LE BLANC¹

ABSTRACT

In 2001, the INSEE conducted a survey to better understand the homeless population. Since there was no survey frame to allow direct access to homeless persons, the survey principle involved sampling the services they received and questioning the individuals who used those services. Weighting the individual input to the survey proved difficult because a single individual could receive several services within the designated reference period. This article shows how it is possible to apply the weight sharing method to resolve this problem. In this type of survey, a single variable can produce several parameters of interest corresponding to populations varying with time. A set of weights corresponds to each definition of parameters. The article focuses, in particular, on "an average day" and "an average week" weight calculation. Information is also provided on the use data to be collected and the nonresponse adjustment.

KEY WORDS: Weight sharing; Incomplete frame; Homeless persons.

1. INTRODUCTION

In 2001, INSEE conducted a survey to better understand the homeless population. This was the first representative survey of this type in France (A survey of this type was conducted in the United States in 1991 by *Research Triangle Institute* (RTI) in the Washington metropolitan area (RTI 1993)). The survey principle was to reach homeless persons through the services provided to them, specifically, overnight accommodation and meals. Obviously, a person could use one or more of the services of the survey frame during the reference period considered, which creates a problem when weighting the survey's individual data files. In this article, we will show how the weight sharing method can be applied to this problem. In this type of survey, unlike most traditional household surveys, a single variable can produce several parameters of interest corresponding to different population concepts: the ones used most often by practitioners are the "average day" and "average week" parameters. A set of weights corresponds to each definition of parameters. We will provide precise definitions of these concepts and will focus in particular on the practical calculation of the corresponding weights. The article is laid out as follows: we will begin by stating the objectives of the survey, identifying its reference population and describing its sample design. We will then introduce the parameters of interest and derive the estimators of these parameters using the weight sharing method. We will describe the practical application of "average day" and "average week" weight calculations. Lastly, we will discuss practical considerations related to the nonresponse adjustment.

2. "HOMELESS" SURVEY

2.1 Objectives of the Survey

The purpose of the survey conducted by the INSEE in February 2001 was to obtain a better understanding of the "homeless" population. This population is normally defined by default as all persons who do not have a fixed residence. It is a population that is not captured by traditional household surveys conducted by the Institut since such surveys have an accommodation survey frame. Since there was no sampling frame for this population, the survey principle involved reaching the target population through the services provided to persons in difficulty, specifically accommodation and meals. These service are provided at certain times that vary depending on their nature: meals are provided every day at noon and in the evening, while overnight accommodation is provided once a day.

This indirect sampling introduces two biases into the population initially targeted and the population actually surveyed. First, the entire target population is not surveyed: only those members who use the services in the survey field are potentially sampled. Second, the population actually surveyed contains individuals who do not belong to the population initially targeted to the extent that the services provided primarily for homeless persons are also used by persons who live in a regular household but who are in a vulnerable situation (this is especially true in the case of meals). Throughout this article, while keeping this distinction in mind, we will however sometimes use the expression "homeless" to designate the persons using the services in the survey field.

¹ Pascal Ardilly and David Le Blanc, Institut National de la Statistique et des Études Économiques, 18 boulevard Adolphe Pinard, 75675, Paris, Cedex, France.
E-Mail : pascal.ardilly@insee.fr, leblanc@ensae.fr.

2.2 Reference Population

The main feature of the services surveyed is that they are provided in specific locations; this location is accordingly called a *centre*. Several types of services correspond to a given centre. The statistical unit sampled, which we will call a *service*, will be defined as a quadruplet (service, day, time interval, person): it consists of a given type of service in a given centre, on a given day, in a given interval of time, to a given person. Of course, a person could receive several services on the same day, let alone in a given week or during the survey month.

The survey *reference period* covers one month (January 15 to February 15, 2001). The total number of days in the survey reference period is designated as J , denoted by the index j .

The *geographic field* of the survey is that of population centres with more than 20,000 inhabitants.

The *services in the survey field* are those that are provided by one of the two types of services retained - meals and accommodation - when they are provided at least one day during the survey reference period.

The *reference population*, designated as $P(J)$, consists of persons who receive at least one service in the survey field during the reference period.

This population of interest depends fundamentally on the reference period. Its size increases with the length of that period, but "more slowly" than the time: in actual fact, certain people are found in the centres every day. In reality, the change in $P(J)$ in relation to J is complex because there are two separate phenomena coming into play that would appear to have different characteristic times:

- at any given time, the "homeless population" only occasionally visits the centres in the frame: to claim to cover that population, it would be necessary to survey over a period of time that would ensure that all persons in this population had used the services at least once (this period is not known but it is acknowledged in France, "according to the experts", that the population not covered during one full month of winter is negligible).
- the "homeless" population is self-renewing over time. Year to year, there are no doubt numerous persons coming into and going out of this population, linked to demographic change or economic or structural changes in society (persons coming into and going out of vulnerable situations).

The question of how to determine J ultimately comes down to knowing whether interest is mainly in a concept of homeless "at a given moment" (J is relatively short) or a concept of homeless over a long period of time (J relatively long). The approach adopted by the INSEE is a compromise between the two.

2.3 Sample Design

The survey's sample design has three stages: selection of population centres, selection of centres and time intervals, and selection of services.

2.3.1 Selection of Population Centres

The first stage of the sample design consists of selecting the population centres, based on a size criterion defined as a combination of the population of the population centres and the ability to provide services so that they could be identified in the records of associations and of the Ministère de la Santé. This first selection stage was carried out several months before the other two. This screening was necessary because the exhaustive census of the centres and the data related to them (type of service provided, average capacity, days open, ...) was then carried out in the selected population centres. This operation was done twice: a detailed survey the year before the data collection and an update just before the start of the data collection. This process produced a survey frame of centres. This frame has a fundamental role: persons who used only non-identified centres were not be sampled.

2.3.2 Selection of Centres, Days and Time Intervals

For practical reasons, it was not possible to survey all of the centres and to keep an interviewer on site at a given centre the entire day. Nor was it possible to interview everyone in a centre. It was therefore imperative to sample:

- centres in the selected population centres (index c)
- survey days during the collection period (index j)
- intervals of time during the survey days (index t).
- persons within one of the selections (centre, day, time interval).

For theoretical reasons, *time intervals were defined in such a way that an individual could not receive two different services during a single time interval* (for example, one of these time intervals was the period from 11:00 a.m. to 2:00 p.m.). It was not reasonably possible to measure the links to the survey frame unless the persons interviewed could easily identify in time and space the services they received during the survey period. In the case of centres offering meals, one time interval covered the noon meal and one time interval covered the evening meal. It was assumed that an individual could use only one centre during the time interval corresponding to the noon meal, otherwise it would be necessary to ask the individual if he had already received a meal somewhere else or if he had eaten twice in the same centre. It was also determined that the length of an interval ensuring use of only one service was also the length of time that an interviewer could reasonably be asked to remain on site interviewing (two to three hours maximum). (Note that daytime accommodation is not part of the services included in the survey field. This restriction of the field reflects two concerns. First, it would be very difficult to divide the day into time intervals of

three or four hours and to determine the links using this breakdown (the memory effort required of the person interviewed would be significant and did not seem reasonable to the survey's designers). Second, it is very difficult to predict the use of these services. We wanted to avoid having a team of interviewers go to a site and not be able to conduct any interviews because of lack of use.)

In actual fact, there is no fundamental difference between the sampling of the centres and the sampling of the periods of time: the relevant units to be considered are the triplets (c, j, t) that correspond to the overlap between a centre, a day and a time interval. Some of the boxes in the "time" and "centres" cross-tabulation table can be eliminated automatically prior to the selection, either because the centre is closed during the time slot considered, or because there is clearly not enough use. (In the latter case, caution must be exercised with respect to the possible restriction of the field should it be found that persons use only this centre and only attend during this time slot. If the latter are atypical, biases will be introduced into the estimations.)

The selection method used was a random selection of the triplets (centres, days, time intervals) in proportion to the size of the centres obtained during the centre census. (In practice, in order to avoid difficulties with centre officials, time intervals were grouped together when a centre was sampled more than four times during the survey period.) Centres were stratified by type. (For accommodation services, centres were stratified by the criteria of men only/women only/mixed accommodation.) However, since this "precautionary" stratification does not apply directly to the observation units, it is useful only if the behaviour of the individuals differs significantly by the type of centre in which they are found.

2.3.3 Selection of Services

This last stage of the sample design consisted in completing the sampling of services, that is, in selecting individuals in a selected centre on a given day during a given time interval. The data collected during the census of the centres were not generally enough to constitute a survey frame of services. Some accommodation centres had lists: this was the more positive scenario where persons could be selected using these lists. However, at the majority of centres (for example, a soup kitchen), it was not even known how many people would show up in a given time interval: it was therefore not possible to develop a survey frame of services. Sampling of the services was done on an equal probabilities basis. As is traditional in multiple stage surveys, selecting a constant number of services (last stage) ensures constant probabilities of selection and thereby limits the risk of expanding sampling variances.

In practice, the selection method used varies from one type of centre to the next, depending on the topography of the sites; existing list, waiting list, arrivals spaced over time, population "grouped" in no order at a single site at the same time, *etc.* It also takes into consideration the

maximum number of interviews that can reasonably be done by the interviewer or interviewers during the survey's time interval, and the fact that it is not desirable to keep the sampled persons too long after the closing of a centre or after meal service has stopped because of the risk of increasing the nonresponse rate.

In all instances, a "counter" counts the number N of services provided during the sampling period. This is crucial to determining the selection probability of the sampled services. At the same time, the counter carries out a standard systematic selection (ideally, the selection should be done by another person (or "sampler") to avoid measurement errors in the use. For budget reasons, it was not possible to resolve this problem) using the following method:

- in centres where a list was available, n services were selected, n being set before the survey;
- in centres without a list, services were selected with a fixed f sampling ratio. f is determined based on the number of expected services \tilde{N} and the number of services that we wanted to sample \tilde{n} in order to ensure equal selection probabilities. In these cases, the size of the sample was not known in advance.

3. PARAMETERS OF INTEREST

The quantities of interest are essentially totals or ratios. We want to estimate a total in relation to a variable y defined for the population $P(J)$,

$$Y_J = \sum_{k \in P(J)} y_k. \quad (1)$$

One specific example of these totals is the size of $P(J)$, $N_J = \text{card}(P(J)) = \sum_{k \in P(J)} 1$.

We also want to estimate the average of y in the reference population,

$$\bar{Y}_J = \frac{Y_J}{N_J} = \frac{1}{N_J} \sum_{k \in P(J)} y_k. \quad (2)$$

For example, y can be the nationality of the individual, the age at which he completed his education, or the number of centres that he visited the day of the interview.

We then have to distinguish between two types of variables:

- variables that are fixed during the survey reference period (such as, age at time of completion of education);
- variables that vary during the survey reference period ($y_k = y_k(j)$). The number of centres visited on the day of the survey fall into this category.

We will begin with the variables that are fixed during the survey reference period. Section 6 looks briefly at those variables that change during that period.

4. ESTIMATION OF A TOTAL OR RATIO IN CASES WHERE THE VARIABLE OF INTEREST IS CONSTANT DURING THE SURVEY PERIOD

For the convenience of the discussion, we will not present explicitly all of the selection stages. Instead, we will use as an example a population centre sampled at the first selection stage.

We note:

- C : all centres in the population centre open at least one day during the survey period, denoted by index c
- $\Pi_{c,j,t}$: all services provided in centre c on day j during time interval t , denoted by index i .
- $\Pi_{j,t}$: all services provided in the population centre on day j during time interval t .
- $P_{c,j,t}$: all persons who visit centre c on day j during time interval t , denoted by index k .
- $P_{j,t}$: all persons who visit a centre in the population centre on day j during time interval t .

Based on the definition of the time intervals, we find that for each individual $k \in P_{j,t}$, there is one and only one service i . Thus, there is a one-to-one correspondence between $P_{j,t}$ and $\Pi_{j,t}$. In other words, for every couple (j, t) , the $P_{c,j,t}$ and $\Pi_{j,t}$ are separate. On the other hand, $P_{c,j,t}$ and P_{c^*,j^*,t^*} can have a non-empty intersection, when $t \neq t^*$.

The population of interest is therefore written

$$P(J) = \bigcup_{c,j,t} P_{c,j,t} = \bigcup_{c \in C} \left(\prod_{c \in C} P_{c,j,t} \right).$$

The central point of the reasoning consists in expressing the total of one variable of the population of *individuals* (which is our total of interest) as the total of another variable of the population of *services* (which are the sampled units), since estimation of the latter does not pose any particular problem. To obtain this result, we can use direct reasoning or apply the weight sharing method, either of which may seem more natural.

Using direct reasoning, we define the application K , which links to each service i received during reference period J in all of the centres in the survey frame the individual who received that service.

$$K : \{\text{services}\} \rightarrow \{\text{individuals}\} \\ i \rightarrow K(i)$$

The population of interest $P(J)$ is represented by K of $\Pi(J)$, all services provided during the reference period in

all centres in the survey field. For each $k \in P(J)$, we define $r_k(J) = \text{card}(K^{-1}(k))$, the number of services provided to individual k during period J in all centres in the survey field, which we will also call the “number of links”.

This gives us the fundamental equation:

$$Y_J = \sum_{k \in P(J)} y_k = \sum_{i \in \Pi(J)} \frac{y_{K(i)}}{r_{K(i)}(J)}. \quad (3)$$

Since variable y takes the same value for all services i “pointing” to individual k , such that $K(i) = k$, the right-hand side can be written

$$\sum_{k \in P(J)} \left[\sum_{i \in \Pi(J); K(i)=k} \frac{y_k}{r_k(J)} \right] = \sum_{k \in P(J)} \frac{y_k}{r_k(J)} \left[\sum_{i \in \Pi(J); K(i)=k} 1 \right].$$

But the quantity in the square brackets is the number of services provided to individual k during period J , or $r_k(J)$, which proves the equation.

We can then see $y_{K(i)}$ as attached to corresponding service i and write y_i in place of $y_{K(i)}$, and $r_i(J)$ in place of $r_{K(i)}(J)$. By using $z_i = y_i/r_i(J)$, $Z = \sum_{i \in \Pi(J)} z_i$, we get $Z = Y_J$.

Formula (3) is none other than the weight sharing formula. The above reasoning is actually the reasoning underlying this method. (Only the expressions change; the weight sharing method describes the links between the sampled population and the population of interest by a matrix rather than an application, a single unit of the sampled population being able to “point” to several units of the population of interest.) The principle of this latter method is set out in Appendix 1.

4.1 Estimation of a Total

Let us now assume that we have a sample s_Π of services to which a set of weights is linked $(w_i)_{i \in s_\Pi}$. We assume these weights are unbiased (this is the inverse of the probabilities of inclusion of services in the sample). s_Π implicitly defines a sample of individuals s_P , which is actually all of the individuals who receive the sampled services. The weight sharing formula (see Appendix 1) ensures that the estimator

$$\hat{Y}_J = \sum_{s_P} y_k \tilde{w}_k$$

is unbiased, where we write for every $k \in s_P$:

$$\tilde{w}_k = \frac{1}{r_k(J)} \sum_{s_\Pi; K(i)=k} w_i. \quad (4)$$

Formula (4) simply states that an individual’s weight is equal to the sum of the weights of the services that were used to “catch him”, divided by the number of links with the survey frame, $r_k(J)$. In this way, it is possible to *work directly on the individuals sampled*: for each individual k , we calculate the weight \tilde{w}_k , and we estimate the total Y_J by \hat{Y}_J .

Figure 1 gives a fictitious sampling example. The service universe contains 13 services, provided to 8 persons. 6 services are sampled. The sample of individuals contains 5 persons, individual number 2 having been "caught" by two different services. Using formula (4), the weights of the individuals sampled will be equal to:

$$\tilde{w}_1 = w_1, \tilde{w}_2 = \frac{1}{2}(w_2 + w_8), \tilde{w}_3 = w_{10}, \tilde{w}_6 = w_7, \tilde{w}_7 = \frac{1}{3}w_9.$$

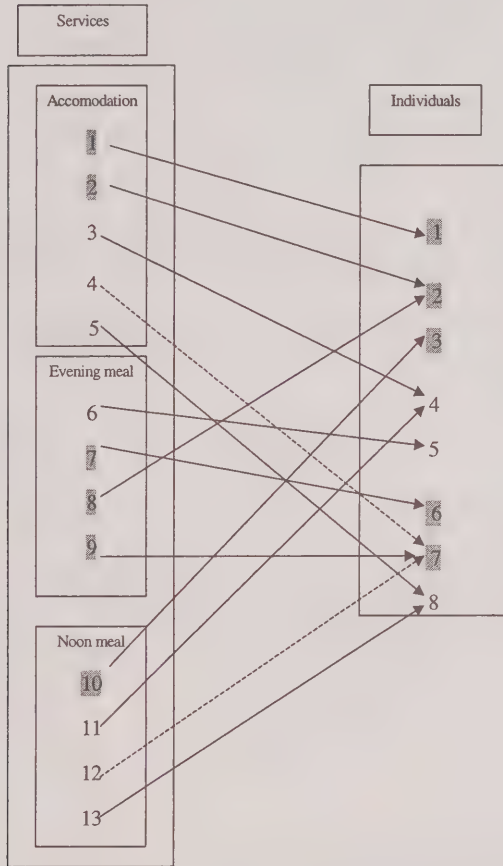


Figure 1. The arrows represent the links between the services and the individuals. The shaded services were sampled. They point to shaded individuals. Dotted lines represent the links reported by individual 7, which were not used to include the individual in the sample.

If the services all have the same weight equal to 13/6 (for example, if the services had been selected by simple random sampling), the number of persons having used services during the survey is estimated by:

$$\hat{Y}_J = \sum_{s_p} \tilde{w}_k = \frac{13}{6} \left[1 + \frac{1}{2} \cdot 2 + 1 + 1 + \frac{1}{3} \right] = \frac{169}{18} \approx 9.39.$$

In this case where the variable being considered does not vary during the survey period, identifying the persons using the services does not affect the estimator bias. Consider an individual "caught" by two different services with weights w_1 and w_2 . In practice, this could produce two cases:

- it is determined that this is the same individual; the weighting associated with this individual will be equal to $(w_1 + w_2)/r_{k(J)}$, and the expression corresponding to the individual in the estimator will be equal to $y_k (w_1 + w_2)/r_{k(J)}$.
- it is not determined that the individual has already been interviewed: two different individuals are counted; the weights associated with these individuals will be equal to $w_1/r_{k(J)}$ and $w_2/r_{k(J)}$, and the expression corresponding to these two pseudo-individuals in the estimator will still be equal to $y_k (w_1 + w_2)/r_{k(J)}$.

Of course, this presumes that the information provided by the same person surveyed in two different locations/on two different days is the same, which is far from given.

However, identifying individuals can be important in order to limit nonresponse (see section 7).

4.2 Estimation of a Ratio

Let us now suppose that we are interested in the estimation of the average \bar{Y}_J (see Formula (2)). \bar{Y}_J can be estimated by the Hajek estimator,

$$\hat{\bar{Y}}_J = \frac{\hat{Y}_J}{\hat{N}_J}$$

where $\hat{N}_J = \sum_{k \in s_p} \tilde{w}_k$.

4.3 Variance Calculation

The variance of the estimators presented above is calculated in the classic manner provided that the reasoning is based on services. The calculation is still complex because it is a multi-stage design with unequal probabilities. To avoid underestimating the true variance, it is essential that all services be retained in cases where several sampled services point to a single individual.

4.4 Comparison with Other Estimating Methods

Having introduced "weight sharing" estimators, it is appropriate to consider an alternative estimating method where we will try to estimate directly the selection probabilities of individuals in the sample. (The weight sharing estimator is not a classic Horvitz-Thompson estimator: the weights of that estimator clearly depend on the complete service sample (see formula (4)). This method can appear more natural. However, we must make two comments:

- it is not reasonably possible to obtain the selection probabilities of physical persons without relying on the services that the individual receives, based on the information provided by the latter when visiting the various centres. Based on the previous expressions, we get:

$$\text{Prob}(k \in s_p) = \text{Prob}\left(\bigcup_{i \in \Pi(J); K(i)=k} i\right)$$

The Poincaré formula enables us to express this probability from single, double, triple, etc probabilities of inclusion of services. Except for the single inclusion probabilities, these are complex probabilities derived as they are from selections of unequal and without replacement probabilities. We cannot therefore hope to obtain a calculable expression for $\text{Prob}(k \in s_p)$. In contrast, the weight sharing method is very simple to apply:

- in a more structured manner, a problem comes from the fact that the selection probabilities of unsampled services are not known in advance because of the multi-stage sample. At the earlier stages, the selection probabilities depend on the previous selection. In our case, we do not know the use of the centres that are not surveyed. To obtain the selection probability of an individual, we must know the inclusion probabilities of *all* services that the individual receives. On the other hand, one of the strengths of the weight sharing method is that the weights of units obtained indirectly (in this case individuals) can only depend on the weights of units sampled directly (services). Lavallée (1995) points out this advantage of the method.

5. ESTIMATION DIFFICULTIES AND PRACTICAL SOLUTIONS IN THE CASE OF A CONSTANT VARIABLE

In the formulae that we have presented, knowing the links between individuals and the services universe is critical. However, these quantities are not known for several reasons:

- a theoretical reason: because the data collection is spread over time, and an individual interviewed at the start of the period cannot anticipate the services that he will use after the interview date (Note that data collection must necessarily be spread over time to ensure good coverage of the target population; synchronous collection, even if technically possible, would not capture the whole target population but only the persons using the services on that date);
- practical reasons: because the memory of the person interviewed becomes questionable after a few days, and because detection by the interviewer or the designer of the survey of the services provided in centres not belonging to the survey frame is very difficult.

In practice, it is therefore impossible to estimate without bias a total of interest over the period of the survey (one month) without making assumptions at the outset (see Section 5.3).

5.1 “Average Day” and “Average Week” estimations

This forces us to look at quantities that bring into play links over a short period, for example, a day or week. The population of persons who use the services in the survey field on a given day j is $P_j = \bigcup_{c,i} P_{c,j,i}$. Let us now introduce the following quantities that relate to day j :

$$\Theta_j = \sum_{k \in P_j} y_k$$

$$N_j = \sum_{k \in P_j} 1 = \text{card}(P_j).$$

If $\tau = \text{card}(J)$ is the number of days in the survey reference period, we define the following parameters of interest:

- the total of y in the population of persons who use the services in the survey field on an “average” day, as follows:

$$\Theta = \frac{1}{\tau} \sum_{j=1}^{\tau} \Theta_j. \quad (5)$$

A specific case is the number of persons who use the services in the survey field on an “average” day, $\bar{N} = 1/\tau \sum_{j=1}^{\tau} N_j$.

In the same way, the average of y in the population of persons who use the services in the survey field on an “average” day is defined as:

$$\psi = \frac{\Theta}{\bar{N}} = \frac{\sum_{j=1}^{\tau} \Theta_j}{\sum_{j=1}^{\tau} N_j}. \quad (6)$$

Defining totals or averages for a given week or an “average week” follows the same principle.

We can estimate these parameters by simply adapting the formulae in the previous section, noting that the $r_k(J)$ must be replaced by the number of services in the survey field that the person sampled received on the day (or week) of the survey.

Note that s_j is the sample of persons interviewed on day j , $r_k(j)$ the number of services in the universe received by individual k on day j only, and $s_k(j)$ the services sampled on day j that link to individual k .

$$\Theta_j \text{ will be estimated by } \hat{\Theta}_j = \sum_{k \in s_j} y_k \tilde{w}_k,$$

$$\text{where } \tilde{w}_k = \frac{1}{r_k(j)} \sum_{i \in s_k(j)} w_i.$$

Here, the weights of the individuals depend on the day j . (But *not* the weights of the services, w_i , which are set one time for all (if there are no nonresponses, this would be the inverse of the selection probabilities of services)). The following analogy is useful to convince oneself of the difference between Θ et Y_j : Consider a service window where everyone who comes must fill out a file. Y_j corresponds to an approach where the person fills out a file the first time that he arrives at the window and does not fill one out on subsequent visits; the “average day” case corresponds to an approach where everyone who arrives at the window fills out a file, regardless of whether he has come to the window on some other day or not. At the end of a week, for example, the analysis of the characteristics of the persons who filled out the files will be very different in the two cases: in the second case, *persons who come to the window often will be over-represented compared to the first case*. It is possible to formalize this approach. We refer interested readers to Ardilly and Le Blanc (1999).

5.2 Practical Estimation of the Links with the Survey Frame

Even if we restrict ourselves to estimating “average week” and “average day” quantities, it is not generally possible to determine the links with the survey frame on a given day (much less a given week or over the whole of the survey period).

5.2.1 “Average Day” Estimation

To share the weights, we must estimate the links relating to the survey day; the situation that presents the most problems is that of persons interviewed at noon in a centre that provides meals; we do not know which centres (meals and/or accommodation) these persons will use that same evening. One option not retained by the INSEE survey designers is to include in the questionnaire questions of the type “Where will you eat (or sleep) this evening?”. The answers can be used to determine the links. Of course the issue is whether the answers to these questions reflect the true links and whether the nonresponse rate for the question would be too high. From a more statistical standpoint, (hypothesizing that there is a certain regularity of behaviour) we could use information relating to the same time interval on the day before the survey. The corresponding links are undoubtedly reasonable approximations of the actual links. The practical problem relates to the possible difference in use of the centres depending on the day of week: for example, some centres are not open on weekends and others are open only on specific days.

5.2.2 “Average Week” Estimation

To share the weights, we retain all the links relating to the week. Clearly, the first option described in 5.2.1 cannot be used. For a given week estimations, we can use, as an approximation of the services used on day j following the interview date, the services used by the individual on day

$(j - 7)$. This is consistent if we assume that there is a certain pattern to the services used depending on the day of the week. This approach would mean that the calendar week would be replaced in estimators by a sliding week, that is, the last seven days beginning on the date of the interview. This is the option that was used for the survey, the questionnaire having been designed to collect the links over the 7 days preceding the interview.

5.3 Estimation Over the Whole of the Survey Period

It may seem that estimating totals and averages for the population $P(J)$ is one of the survey’s objectives. This estimation calls on the links between individuals sampled and the services in the survey field during the whole of the data collection period, which are not known. This means that we have to model the evolution of the links beyond a week or, what amounts to the same thing, model the use behaviour of the individuals in the centres.

The solution is not simple. For example, the hypothesis that comes to mind is

$$\forall k, r_k(J) = A \cdot r_k(S) \quad (7)$$

where A is the number of weeks of the survey and $r_k(S)$ is the number of links for individual k with the services of the survey field during a week S , leads to estimators for the whole of the period that are identical to the estimators for an average week. In effect, an “average week” estimator weights individual k by

$$\sum_{i \in S_k(J)} \frac{w_i}{A \cdot r_k(S_i)}$$

where S_i is the week during which he received service i and $s_k(J)$ is the sampling of services that link to individual k , whereas a theoretical “whole period” estimator weights the individual k by

$$\sum_{i \in s_k(J)} \frac{w_i}{r_k(J)}.$$

Equation (7) is therefore an adequate condition of equality of these estimators. This condition is satisfied in particular when for any j and any k

$$r_k(J) = \text{card}(J) \cdot r_k(j) \quad (8)$$

that is, when the number of daily links does not depend on j .

This hypothesis is definitely too strong. To expand on this point, we will have to use the data provided by the survey itself on the behaviour of the individuals with respect to use of the centres.

The most sought after figure of the survey – in the French context – is undoubtedly an estimate of the size of the “homeless” population, that is, an estimation of the size of $P(J)$. In addition to the issues regarding counting the links that have already been discussed extensively, this estimation runs up against several inadequacies in the survey frame as well as the indirect nature of the sampling.

- The risk of overlooking certain structures when identifying the centres is significant. Even with an exhaustive inventory, the gap between when the inventory is established and the survey itself takes place makes it likely that new unidentified centres will appear in the survey frame. This can introduce a bias to the extent that some individuals who might use these structures would not use any other service in the survey frame. (We might also expect those in charge of certain centres to refuse to cooperate: for the INSEE survey, there was virtually no refusal by the institutions (less than 1% refusal rate). This was due largely to consideration awareness building at the time the centres were identified and just before the survey.) Further, the lack of bias depends on a correct calculation of the links; use of centres not included in the frame should not be counted in these links.
- Individuals who use the centres only outside the “classic” hours (those in which we have the means to count the services) are outside the survey frame. (Counting them would create significant on-site implementation problems.)
- Another source of bias can come from the careful counting of the total number of services provided in the centres during the survey, these numbers being used to calculate the probability of a service being sampled. For budget reasons, one person only counted the services and did the sampling, a situation that could create problems of rigour in the sampling if there is confusion in the field.
- In terms of the concepts, the only remaining problem was that the survey had to take place over a month and that the target population may have changed during that period.

The estimation of the size of the population is therefore particularly fragile. For this reason, we can expect any errors to be larger for the totals than for the averages.

6. ESTIMATION IN THE CASE OF VARIABLES OF INTEREST THAT ARE NOT CONSTANT OVER THE SURVEY PERIOD

Some of the survey’s variables of interest depend on the observation date and therefore are not constant over the survey period. This can be the case with answers to questions dealing with the day before the interview, for example “How many meals did you have yesterday?”, “How many times did you sleep in the street last week?”, etc. The questions on links also fall into this category. It is therefore important to determine the extent to which we can adapt the earlier formalism to estimations involving this type of variable. In other words, where y is such a variable of interest.

If we go back to expression (3), it is easy to see that the constancy of y_k during the survey period is the condition

that makes it possible to factor y_k and to reveal the links $r_k(J)$. From this we can deduce that *the above type of calculation is always valid for estimations covering shorter periods than the period for which the y_k are constant.*

This means that for variables that are constant for a day, we can appropriately use the “average day” estimators. For variables that are constant over the week, we can use the “average day” or “average week” estimators.

7. ADJUSTMENT FOR TOTAL NONRESPONSE

To describe the operation fully, we still need to explain how to move from a set of inclusion probabilities (and thus initial weights of services included in the sample) to a set of weights on respondent services. Some people will agree to the interview, others will not. We will refer to services in the first case as respondent services and those in the second case as nonrespondent services. The usual adjustment methods for total nonresponse can be applied. We suggest a nonresponse adjustment by homogeneous subgroup (for a description of the method, see for example Hambaz and Legendre 1999).

In reality, the main problem relates to the fact that there is no survey frame of individuals and thus no advance information on nonrespondents. In a world that is likely very heterogeneous, this is a considerable handicap. We therefore have to model the service response behaviour. We know from the test surveys of the INED (Institut National des Etudes Démographiques) that nonresponse varies widely depending on the type of centre (Firdion and Marpsat 1997). Other variables in the survey frame can be used to build homogeneous groups (day of the week, period of the day, groups of population centres, ...).

A reweighting of the respondent services produces weights for the respondent services of the type

$$w_i = 1/\delta_i\pi_i, \text{ where}$$

π_i is the probability of inclusion of service i in the sample

δ_i is the probability estimated after the fact that service i will result in a response.

This provides us with a set of weights for the respondent services.

In fact, some of the nonresponses come from the fact that the same individual is sampled several times: obviously, an individual who is sampled twice might respond the first time but not the second. (The frequency of occurrence of this event was not known at the time of writing this paper.) The second selection therefore produces a “false non-response”. If this is not detected, the total nonresponse adjustment procedure leads to an incorrect reweighting, when the true value can be obtained from a questionnaire that has already been completed. To avoid this problem, the interviewer tries to find out the reason for the refusal and must check off a specific box when the individual states that he has already been interviewed. In this situation, the

interviewer collects some information, including the first name and the date of birth, that can be used to link this questionnaire to the questionnaire that has already been completed. (The ideal situation would be to have an identifier for the respondents. This approach was not used because of confidentiality requirements and consideration of the reaction of the persons interviewed to such a measure.) However, in the field, it can be difficult to obtain a reason for refusal. Even if a reason is given, problems can occur. (It is hard to verify that a person who states that he or she has already been interviewed has in fact been interviewed. Even if the person is showing goodwill, he may have been interviewed a few days earlier for a completely different survey than the INSEE survey.)

8. CONCLUSION

In this article, we show how the weight sharing method can be used to weight the survey conducted by the INSEE in order to better understand homeless persons. The method has many advantages. It makes it possible to work on a file of individuals, that is, on the natural statistical units used in the definition of the parameters of interest. Simple to apply, it also makes it easy to move from one reference period to another ("average day", "average week" estimation). Operations following to the survey, such as the nonresponse adjustment and the calculation of variance can be carried out in a traditional framework because they are done on sampled units (services), for which the selection probabilities are known, and not on individuals, for which the selection probabilities are not known. We show that a crucial quality criterion of such a survey is reliable data collection on use of services by the persons interviewed. Without these data, it is not possible to weight the survey. The weight sharing method appears to be a good compromise for a survey in which the purpose is not simply to count a population but to better understand it through the use of a questionnaire. Other alternative methodologies could be used for a survey aimed simply at determining the size of the homeless population. The first such methodology uses capture-recapture techniques to determine the size of animal populations (see for example, Pollock, Turner and Brown 1994). These techniques cannot be easily applied to a population that is often suspicious of any attempt to identify it, which they perceive negatively. Another technique is that of "snowball" sampling, which involves finding individuals of interest through the intermediary of individuals already sampled (Franck and Snijders 1994). It relies on a system of mutual knowledge of persons, who are probably illusive in the community. These methods always run up against the issue of the identifying individuals. In our case, the only places where it is possible to find the persons we are seeking are the centres: it is essential that we work through the centres.

ACKNOWLEDGEMENTS

The authors thank the journal's Editor and two anonymous referees whose comments helped improve both the content and layout of the article. Any errors that remain are entirely our responsibility.

APPENDIX 1: THE WEIGHT SHARING METHOD APPLIED TO THE PROBLEM

This appendix briefly presents the principle of the weight sharing method. For a more complete discussion, the reader may consult Lavallée (1995) or Deville (1999) whose notations we have used.

1. We have a population U of n units, and a population V of m units. The units of U are services in the survey field. The units of V are persons who used at least one service during the survey period (otherwise expressed in the present case as $V = P(J)$ with the previous notations).
2. It is assumed that there are links between the units of the two populations. These links can be written in the form of a matrix

$$(r_{ik}) \quad 1 \leq i \leq n, \\ 1 \leq k \leq m$$

where $r_{ik} = 1$ if unit k of V is linked to unit i of U , $r_{ik} = 0$ otherwise. In this case, the links connect the services to the persons who used these services: $r_{ik} = 1$ if person k used service i of U , $r_{ik} = 0$ otherwise.

3. All units of U have at least one link to a unit of V . Clearly, that is achieved here by definition of population V . Further, in this case, each unit of population U points to one and only one unit of V .

In general, we are interested in the total of a variable of interest y in V ,

$$Y = \sum_{k \in V} y_k.$$

If, for example, we use $y = 1$, the total of interest is the number of persons who used a service in the survey field during the month of the survey.

We can write

$$r_k = \sum_{i \in U} r_{ik}.$$

The identity $Y = \sum_{i \in U} \sum_{k \in V} (r_{ik}/r_k) y_k$ makes it possible to define for any $i \in U$ the variable $z_i = \sum_{k \in V} (r_{ik}/r_k) y_k$ which gives:

$$Z = \sum_{i \in U} z_i = \sum_{k \in V} y_k = Y.$$

Let us now assume that we have a sample s_U from the population U , which is associated with a set of weights $(w_i)_{i \in s_U}$. This sample implicitly defines a sample in V , s_V , specifically

$$s_V = \{k \in V; \exists i \in s_U, r_{ik} = 1\}.$$

We assume that we collected the r_{ik} for all $k \in s_V$, that is, that all links between individuals and the universe U are known (this point is fundamental).

The total $Z = Y$ is estimated by $\hat{Z} = \sum_{s_U} w_i z_i$.

And consequently, if the weights are unbiased (that is, set so that \hat{Z} is without bias), \hat{Y} estimates Y without bias.

We can rewrite $\hat{Z} = \sum_{s_U} w_i \sum_{k \in V} r_{ik} y_k / r_k = \hat{Y}$.

The second equation impacts only s_V by definition and therefore $\hat{Y} = \sum_{s_V} y_k (\sum_{s_U} w_i r_{ik} / r_k) = \sum_{s_V} y_k \tilde{w}_k$, where we have written for all $k \in s_V$:

$$\tilde{w}_k = \frac{1}{r_k} \sum_{s_U} w_i r_{ik}. \quad (9)$$

We can work directly on the individuals sampled. In our case, r_k is the number of links, that is, the number of services used by the person interviewed during the survey reference period. It is the quantity that is written $r_k(J)$ in the previous sections, the dependence on J being intended to remind that links affecting the weight can vary by the type of estimator ("average day", "average week") considered. This number is derived from the use data collected in the survey.

APPENDIX 2: SUMMARY TABLE OF EXPRESSIONS

J	All days in the survey reference period
τ	= card(J), number of days in the reference period
$P(J)$	population of interest, all persons who used at least one service in the survey field during the reference period
N_j	= card($P(J)$), size of the population of interest
C	all centres in the population centre, denoted by index c
$\Pi_{c,j,t}$	all services provided in centre c on day j during time interval t , denoted by index i
$\Pi_{j,t}$	all services provided in the population centre on day j during time interval t
$P_{c,j,t}$	all persons who visit centre c on day j during time interval t , denoted by index k
$P_{j,t}$	all persons who visit one of the centres in the population centre on day j during time interval t
P_j	all persons who use services in the survey field on day j
y	variable of interest
Y_J	total of variable y in the reference population
\bar{Y}_J	average of y in the reference population

$\Pi(J)$	all services provided during the reference period in all centres in the survey frame
$r_k(J)$	number of services provided to individual k during period J in all centres in the survey field, or "number of links"
s_Π	sample of services
w_i	weight associated with the services sample
s_P	sample of individuals, all individuals who received sampled services
\tilde{w}_k	weight associated with the sample of individuals
Θ_j	total of y in P_j
N_j	= card(P_j)
Θ	total of y "an average day"
\bar{N}	number of persons on "an average day"
Ψ	= $\frac{\Theta}{\bar{N}}$, average of y "on an average day"
$r_k(j)$	number of services received by the individual k on day j only
s_j	sample of persons interviewed on day j
$s_k(j)$	all services sampled on day j that point to individual k
$s_k(J)$	all services sampled during period J that point to individual k

REFERENCES

- ARDILLY, P., and LE BLANC, D. (1999). Enquête auprès des personnes sans-domicile : éléments techniques sur l'échantillonnage et le calcul de pondérations individuelles, une application de la méthode du partage des poids. Working Paper, INSEE, F9903.
- CHAMBAZ, C., and LEGENDRE, N. (1999). Calcul des pondérations dans le panel européen de ménages. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.
- DEVILLE, J. C. (1999). Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? suivi de : comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.
- FIRDION, J. M., and MARPSAT, M. (1997). Comptes rendus du groupe « pondérations » de l'enquête auprès des personnes sans-domicile, mimeo.
- FRANCK, O., and SNIJDERS, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- POLLOCK, K.H., TURNER, S.C. and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117-124.
- RTI (1993). Prevalence of drug use in the Washington DC metropolitan area, homeless and transient population : 1991. *Technical report*, 2.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 16, Number 4, 2000

Borrowing Strength When Explicit Data Pooling Is Prohibited <i>Jerome P. Reiter</i>	295
Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration <i>Johan Bring and Kenneth Carling</i>	321
Income Measurement Error in Surveys: A Review <i>Jeffrey C. Moore, Linda L. Stinson, and Edward J. Welniak, Jr.</i>	331
On Variance Estimation for Measures of Change When Samples Are Coordinated by the Use of Permanent Random Numbers <i>Lennart Nordberg</i>	363
A Functional Form Approach to Calibration <i>Victor M. Estevao and Carl-Erik Särndal</i>	379
An Assessment of the Current State of Dependent Interviewing in Household Surveys <i>Nancy A. Mathiowetz and Katherine A. McGonagle</i>	401
Remembering Heads and Bushels: Cognitive Processes Involved in Agricultural Establishments' Reports of Inventories <i>Jaki Stanley McCarthy and Martin A. Safer</i>	419
Accuracy of Using Pneumonia as an Underlying Cause in the Cause-of-Death Register <i>Boo Svartbo, Linda Nilsson, Anders Eriksson, Gösta Bucht, Lars Age Johansson, and Lars Olov Bygren</i>	435
Index to Volume 16, 2000	445

Volume 17, Number 1, 2001

Can A Statistician Deliver? <i>Richard Platek and Carl-Erik Särndal</i>	1
Comment <i>Barbara A. Bailer</i>	21
<i>Paul P. Biemer</i>	25
<i>Alain Desrosières, Jean-Claude Deville, and Olivier Sautory</i>	33
<i>Eva Elvers and Lennart Nordberg</i>	39
<i>Ivan P. Fellegi</i>	43
<i>Robert M. Groves and Nancy A. Mathiowetz</i>	51
<i>David Holt</i>	55
<i>Margarida Madaleno</i>	63
<i>David A. Marker and David R. Morganstein</i>	71
<i>Pilar Martín-Guzmán</i>	73
<i>Photis Nanopoulos</i>	77
<i>Svein Nordbotten</i>	87
<i>Erkki Pahkinen</i>	93
<i>Jaki Stanley McCarthy</i>	99
<i>Lynne Stokes</i>	103
<i>Dennis Trewin</i>	107
Rejoinder <i>Richard Platek and Carl-Erik Särndal</i>	113
Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis <i>Paul P. Biemer, Henry Woltman, David Raglin, and Joan Hill</i>	129
Estimation of Interviewer Effects on Multivariate Binary Responses in a Community Based Survey <i>Sujuan Gao</i>	149
Swedish Employment in the 1950s – How to Fill the Lacuna <i>Gudmundur Gunnarsson and Thomas Lindh</i>	163
Practicing What We Preach: The Application of Continuous Improvement in a Preclinical Statistics Department at a Pharmaceutical Company <i>V. Bill Pikounis, Joseph M. Antonello, Danielle Moore, Edith T. Senderak, and Keith A. Soper</i>	187

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS

TABLE DES MATIÈRES

Volume 29, No. 1, March/mars 2001, 1-172

Richard A. LOCKHART Message from the new Editor/Un message du nouveau rédacteur en chef	1
Christian GENEST: Report from the former Editor/Rapport du rédacteur en chef sortant	3
Malay GHOSH and Yeong-Hwa KIM The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis	5
Daniel R. ENO and Keying YE Probability matching priors for an extended statistical calibration model	19
Patrick E. BROWN and Piet de JONG Nonparametric smoothing using state space techniques	37
Kert VIELE Evaluating fit in functional data analysis using model embeddings	51
Thomas J. DiCICCIO, Michael A. MARTIN and Steven E. STERN Simple and accurate one-sided inference from signed roots of likelihood ratios	67
Daniel B. HALL On the application of extended quasi-likelihood to the clustered data case	77
Paul KABAILA and John BYRNE Exact short Poisson confidence intervals	99
Vincent F. MELFI, Connie PAGE and Margarida GERALDES An adaptive randomized design with application to estimation	107
Giseon HEO, Byron SCHMULAND and Douglas P. WIENS Restricted minimax robust designs for misspecified regression models	117
Min TSAO and Julie ZHOU On the robustness of empirical likelihood ratio confidence intervals for location	129
Marc HALLIN, Amal MELLOUK and Khalid RIFI Projection de Hájek et polynômes de Bernstein	141
Faouzi EL BANTLI and Marc HALLIN Asymptotic behaviour of M-estimators in AR(p) models under nonstandard conditions	155
Acknowledgement of referees' services/Remerciements aux membres des jurys	169
Corrigenda and addendum	170
Forthcoming Papers/Articles à paraître	171

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préférablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0, I, 1).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Richard A. LOCKHART	1
Message from the new Editor/Un message du nouveau rédacteur en chef	
Christian GENEST:	
Report from the former Editor/Rapport du rédacteur en chef sortant	3
Malay GHOSH and Yeong-Hwa KIM	5
The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis	5
Daniel R. ENO and Keying YE	19
Probability matching priors for an extended statistical calibration model	19
Patrick E. BROWN and Piet de JONG	37
Nonparametric smoothing using state space techniques	37
Kert VIELE	51
Evaluating fit in functional data analysis using model embeddings	51
Thomas J. DICICCIO, Michael A. MARTIN and Steven E. STERN	67
Simple and accurate one-sided inference from signed roots of likelihood ratios	67
Daniel B. HALL	77
On the application of extended quasi-likelihood to the clustered data case	77
Paul KABALLA and John BYRNE	99
Exact short Poisson confidence intervals	99
Vincent F. MEIRI, Connie PAGE and Margarida GERALDES	107
An adaptive randomized design with application to estimation	107
Giseon HEO, Byron SCHMULAND and Douglas P. WIENS	117
Restricted minimax robust designs for misspecified regression models	117
Min TSAO and Julie ZHOU	129
On the robustness of empirical likelihood ratio confidence intervals for location	129
Marc HALLIN, Amal MEILOUK and Khalid RIFI	141
Projection de Hájek et polynômes de Bernstein	141
Faouzi EL BANTLI and Marc HALLIN	155
Asymptotic behaviour of $M(p)$ models under nonstandard conditions	155
Acknowledgement of referees' services/Remerciements aux membres des jurys	169
Corrigenda and addendum	170
Forthcoming Papers/Articles à paraître	171

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 16, Number 4, 2000

Borrowing Strength When Explicit Data Pooling Is Prohibited

Jerome P. Reiter

Attrition and Misclassification of Drop-outs in the Analysis of Unemployment Duration

Johan Brmg and Kenneth Carling

Income Measurement Error in Surveys: A Review

Jeffrey C. Moore, Linda L. Shinson, and Edward J. Welniak, Jr.

On Variance Estimation for Measures of Change When Samples Are Coordinated by the Use of Permanent Random Numbers

Lennart Nordberg

A Functional Form Approach to Calibration

Victor M. Estevao and Carl-Erik Särndal

An Assessment of the Current State of Dependent Interviewing in Household Surveys

Nancy A. Mathiowetz and Katherine A. McGonagle

Remembering Heads and Bushels: Cognitive Processes Involved in Agricultural Establishments' Reports of Inventories

Jaki Stanley McCarthy and Martin A. Sager

Accuracy of Using Pneumonia as an Underlying Cause in the Cause-of-Death Register

Boo Svartholm, Linda Nilsson, Anders Eriksson, Gösta Buch, Lars Åge Johansson, and Lars Olov Bygren

Index to Volume 16, 2000

Volume 17, Number 1, 2001

Can A Statistician Deliver?

Richard Platek and Carl-Erik Särndal

Comment

Barbara A. Bailar

Paul P. Biemer

Alain Desrosières, Jean-Claude Deville, and Olivier Sautory

Eva Elvers and Lennart Nordberg

Ivan P. Fellegi

Robert M. Groves and Nancy A. Mathiowetz

David Holt

Margarida Madaleno

David A. Marker and David R. Morganstein

Pilar Martin-Guzman

Photis Nanopoulos

Svein Nordbotten

Erkki Pakkinen

Jaki Stanley McCarthy

Lynne Stokes

Dennis Trewin

Rejoinder

Richard Platek and Carl-Erik Särndal

Enumeration Accuracy in a Population Census: An Evaluation Using Latent Class Analysis

Paul P. Biemer, Henry Wolman, David Raglin, and Joan Hill

Estimation of Interviewer Effects on Multivariate Binary Responses in a Community Based Survey

Sijuan Gao

Swedish Employment in the 1950s – How to Fill the Lacuna

Gudmundur Gunnarsson and Thomas Lindh

Practicing What We Preach: The Application of Continuous Improvement in a Preclinical Statistics Department

at a Pharmaceutical Company

V. Bill Pitkouris, Joseph M. Antonello, Danielle Moore, Edith T. Sanderak, and Keith A. Soper

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

BIBLIOGRAPHIE

IT(J) ensemble des prestations servies durant la période de référence dans l'ensemble des centres du champ de l'enquête

$r_k(J)$ nombre de prestations servies à l'individu k durant la période J dans l'ensemble des centres du champ de l'enquête, ou "nombre de liens"

s_{π} échantillon de prestations

w_i poids associés à l'échantillon de prestations

s_p échantillon d'individus, ensemble des individus destinataires des prestations échantillonnées

\tilde{w}_k poids associés à l'échantillon d'individus

Θ_j total de y sur P_j

N_j = card (P_j)

Θ total de y "un jour moyen"

\bar{N} nombre de personnes "un jour moyen"

$\psi = \frac{\bar{N}}{\Theta}$, moyenne de y "un jour moyen"

$r_k(J)$ nombre de prestations reçues par l'individu k le jour j uniquement

s_j échantillon des personnes interrogées le jour j

$s_k(J)$ ensemble des prestations échantillonnées le jour j qui renvoie à l'individu k

$s_k(J)$ ensemble des prestations échantillonnées pendant la période J qui renvoie à l'individu k

CHAMBAZ, C., et LEGENDRE, N. (1999). Calcul des pondérations dans le panel européen de ménages. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.

DEVILLE, J. C. (1999). Les enquêtes par panel : en quoi diffèrent-elles des autres enquêtes? suivi de : comment attraper une population en se servant d'une autre. *Actes des journées de méthodologie statistique, INSEE Méthodes*, 84-86.

FIRDION, J. M., et MARPSAT, M. (1997). Comptes rendus du groupe « pondérations » de l'enquête auprès des personnes sans-domicile, mimeo.

FRANCK, O., et SNIJDERS, T. (1994). Estimating the Size of Hidden Populations Using Snowball Sampling. *Journal of Official Statistics*, 10, 53-67.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

POLLOCK, K.H., TURNER, S.C. et BROWN, C.A. (1994). Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête*, 20, 121-128.

RTT (1993). Prevalence of Drug Use in the Washington DC Metropolitan Area, Homeless and Transient Population : 1991. *Technical report*, 2.

ANNEXE 1 :
LA MÉTHODE DU PARTAGE DES
POIDS APPLIQUÉE AU PROBLÈME

Cette annexe rappelle brièvement le principe de la méthode du partage des poids. Pour un exposé plus complet, le lecteur pourra consulter Lavallée (1995), ou Deville (1999) dont nous reprenons les notations.

1. On dispose d'une population U de n unités, et d'une population V de m unités. Ici, les unités de U sont les prestations dans le champ de l'enquête. Les unités de V sont les personnes ayant bénéficié d'au moins une prestation pendant la période de l'enquête (autrement dit dans le cas présent $V = P(J)$ avec les notations précédentes).

2. On suppose qu'il existe des liens entre les unités des deux populations. Ces liens peuvent s'écrire sous la forme d'une matrice

$$(r_{ik}) \quad 1 \leq i \leq n, \quad 1 \leq k \leq m$$

où $r_{ik} = 1$ si l'unité k de V est reliée à l'unité i de U , $r_{ik} = 0$ sinon. Ici, les liens relient les prestations aux personnes ayant fréquenté ces prestations : $r_{ik} = 1$ si la personne k a fréquenté la prestation i de U , sinon.

3. Toutes les unités de U ont au moins un lien avec une unité de V . Cela est évidemment réalisé ici, par définition de V . De plus, ici, chaque unité de la population U pointe sur une unité et une seule de V . Dans le cas général, on s'intéresse au total d'une variable d'intérêt y sur V ,

$$Y = \sum_{k \in V} y_k.$$

Si par exemple on prend $y = 1$, le total d'intérêt est le nombre de personnes ayant fréquenté un service du champ de l'enquête pendant le mois de l'enquête.

On note

$$r_k = \sum_{i \in U} r_{ik}.$$

L'identité $Y = \sum_{i \in U} \sum_{k \in V} (r_{ik}/r_k) y_k$ permet de définir pour tout $i \in U$ la variable $z_i = \sum_{k \in V} (r_{ik}/r_k) y_k$ et on a :

$$Z = \sum_{i \in U} z_i = \sum_{k \in V} y_k = Y.$$

Supposons maintenant que l'on dispose d'un échantillon s_U issu de la population U , auquel est associé un jeu de poids $(w_i)_{i \in s_U}$. Cet échantillon définit implicitement un échantillon dans V , précisément $s_V = \{k \in V, \exists i \in s_U, r_{ik} = 1\}$.

On suppose que l'on a collecté les r_{ik} pour tous $k \in s_V$, c'est-à-dire que tous les liens des individus avec l'univers U sont connus (ce point est fondamental).

Le total $Z = Y$ est estimé par $\hat{Z} = \sum_{i \in s_U} w_i z_i$. Et donc, si les poids sont sans biais (c'est-à-dire, établis de manière que \hat{Z} est sans biais), \hat{Y} estime sans biais Y . On peut réécrire $\hat{Z} = \sum_{i \in s_U} \sum_{k \in V} r_{ik} y_k / r_k = \hat{Y}$. La deuxième somme ne porte que sur s_V par définition, et donc $\hat{Y} = \sum_{i \in s_U} \sum_{k \in s_V} w_i r_{ik} / r_k = \sum_{i \in s_U} w_i \tilde{w}_k$, où l'on a posé pour tout $k \in s_V$:

$$\tilde{w}_k = \frac{1}{\sum_{i \in s_U} w_i r_{ik}}. \quad (9)$$

On peut donc travailler directement sur les individus échantillonnés. Dans notre cas, r_{ik} est le nombre de liens, c'est-à-dire le nombre de services fréquentés par la personne interrogée pendant la période de référence de l'enquête. C'est la quantité qui est notée $r_k(j)$ dans les sections précédentes, cette dépendance en j étant destinée à rappeler que les liens intervenant dans les poids peuvent varier selon le type d'estimateur (« jour moyen », « semaine moyenne ») que l'on considère. Ce nombre se déduit des données de fréquentation collectées à l'enquête.

ANNEXE 2 :
TABLEAU RÉCAPITULATIF DES NOTATIONS

J	ensemble des jours de la période de référence de l'enquête
τ	= card(J), nombre de jours de la période de référence
$P(J)$	population d'intérêt, ensemble des personnes qui ont fréquenté au moins une prestation du champ de l'enquête pendant la période de référence
N_j	= card($P(J)$), effectif de la population d'intérêt
C	ensemble des centres de l'agglomération, repérés par l'indice c
$\Pi_{c,j,t}$	ensemble des prestations servies dans le centre c le jour j pendant l'intervalle de temps t , repérées par l'indice i
$\Pi_{j,t}$	ensemble des prestations servies dans l'agglomération le jour j pendant l'intervalle de temps t
$P_{c,j,t}$	ensemble des personnes se présentant dans le centre c le jour j pendant l'intervalle de temps t , repérées par l'indice k
$P_{j,t}$	ensemble des personnes qui fréquentent les services du champ de l'enquête un jour j
Y	variable d'intérêt
Y_j	total de la variable y dans la population de référence
\bar{Y}_j	moyenne de y dans la population de référence

auparavant pour une toute autre enquête que l'enquête de l'INSEE).

8. CONCLUSION

Dans cet article, nous montrons comment la méthode du partage des poids peut être utilisée pour pondérer l'enquête menée par l'INSEE pour mieux connaître les personnes sans-domicile. La méthode présente un grand nombre d'avantages. Elle permet de travailler sur un fichier d'individus, c'est-à-dire sur les unités statistiques naturelles utilisées dans la définition des paramètres d'intérêt. Simple à mettre en oeuvre, elle permet de passer aisément d'une période de référence à une autre (estimation « un jour moyen », « une semaine moyenne »). Les opérations en aval de l'enquête comme la correction de la non-réponse et le calcul de variance peuvent être réalisées dans un cadre classique, car elles se font sur les unités échantillonnées (les prestations), dont on maîtrise les probabilités de tirage, et non sur les individus, dont les probabilités de tirage sont inconnues. Nous montrons qu'un critère crucial de qualité d'une telle enquête est le recueil fidèle des données de fréquentation des services par les personnes interrogées. Sans ces données, il n'est pas possible de pondérer l'enquête. La méthode du partage des poids paraît un bon compromis pour une enquête dont le but n'est pas seulement de dénombrer une population, mais de mieux la connaître en passant un questionnaire. Pour une enquête visant principalement à un dénombrement des personnes sans-domicile, des méthodologies alternatives pourraient être envisagées. La première tourne autour des techniques de capture-recapture, utilisées pour connaître les effectifs de populations animales (à ce sujet, voir par exemple Pollock, Turner et Brown 1994). Ces techniques ne sont pas aisées à mettre en oeuvre dans une population souvent réticente à toute tentative d'identification, perçue négativement. Une autre technique est celle de l'échantillonnage « boule de neige », qui consiste à aller chercher les individus d'intérêt par l'intermédiaire d'individus déjà échantillonnés (Frank et Snijders 1994). Elle s'appuie sur un système de connaissances mutuelles des personnes vraisemblablement illusoires dans ce milieu. Ces méthodes se heurtent toujours à la question du repérage des individus. Dans notre cas, les seuls lieux où l'on peut trouver les personnes avec une probabilité suffisamment importante sont les centres : passer par l'intermédiaire des centres est incontournable.

REMERCIEMENTS

Les auteurs remercient le rédacteur de la revue ainsi que deux rapporteurs anonymes, dont les remarques ont permis d'améliorer le fond et la forme de l'article. Les erreurs qui demeureront nous sont entièrement imputables.

L'échantillon) à un jeu de poids sur les prestations répondantes. En effet, certaines personnes vont accepter l'entrevue, d'autres non. On parlera dans le premier cas de prestation répondante, dans le deuxième de prestation non répondante. Les méthodes habituelles de correction de la non-réponse totale peuvent être mises en oeuvre. Nous suggérons une correction de la non-réponse par sous-groupes homogènes (pour une description de la méthode, voir par exemple Chambaz et Legendre 1999). Concrètement, la difficulté majeure tient au fait qu'il n'y a pas de base de sondage d'individus, et donc pas d'information *a priori* sur les non-répondants. Dans un monde probablement très hétérogène, c'est un handicap considérable. On modélise donc le comportement de réponse des prestations. On sait depuis les enquêtes expérimentales de l'INED (Institut National des Etudes Démographiques) que la non-réponse varie fortement selon le type de centre (Fridon et Marpsat 1997). D'autres variables de la base de sondage peuvent être utilisées pour constituer des groupes homogènes (jour de la semaine, période du jour, groupes d'agglomérations, ...).

$w_i = 1 / \delta_i \pi_i$, où π_i est la probabilité d'inclusion de la prestation i dans l'échantillon δ_i est la probabilité estimée *a posteriori* que la prestation i donne lieu à réponse.

On obtient ainsi un jeu de poids pour les prestations

répondantes.

En fait, certaines non-réponses viennent du fait qu'un même individu est échantillonné plusieurs fois : on peut penser qu'un individu échantillonné deux fois réponde lors du premier tirage, mais pas lors du second. (La fréquence d'occurrence de cet événement n'était pas connue au moment de la rédaction de cet article). Le second tirage génère alors une « fausse non-réponse ». Si celle-ci n'est pas détectée, la procédure de correction de la non-réponse totale amène à répondre à tort, alors que la vraie valeur peut être récupérée dans un questionnaire déjà rempli. Pour éviter cela, l'enquêteur cherche à connaître le motif des refus et il doit cocher une case spécifique lorsque l'individu déclare avoir déjà été interrogé. Dans ce cas, il collecte quelques informations, dont le prénom et la date de naissance, qui doivent servir à relier ce questionnaire avec un questionnaire déjà rempli. (L'idéal serait de disposer d'un identifiant en compte de l'accueil d'une telle mesure par les personnes interrogées ont conduit à ne pas retenir cette idée). Sur le terrain cependant, il est difficile d'obtenir le motif du refus. Même si on en dispose, des difficultés peuvent subsister. (Il est difficile de vérifier qu'un individu qui déclare avoir déjà été interrogé l'a effectivement déjà été. Même si l'individu est de bonne foi, il peut avoir été interrogé quelques jours

champ de l'enquête pendant une semaine S , conduit à des estimateurs sur l'ensemble de la période identiques aux estimateurs sur une semaine moyenne. En effet, un estimateur « semaine moyenne » pondère l'individu k par

$$\sum_{i \in s_k(S)} \frac{A_i r_k(S)}{w_i}$$

où S_i est la semaine durant laquelle la prestation i lui est servie, et $s_k(S)$ est l'échantillon de prestations qui renvoient à l'individu k , alors qu'un estimateur théorique « ensemble de la période » pondère l'individu k par

$$\sum_{i \in s_k(S)} \frac{r_k(S)}{w_i}$$

La relation 7 est donc une condition suffisante d'égalité de ces estimateurs. Cette condition est notamment satisfaite si pour tout j et tout k

$$(8) \quad r_k(j) = \text{card}(j) \cdot r_k(j)$$

c'est-à-dire si le nombre de liens journaliers ne dépend pas de j .

Cette hypothèse est certainement trop forte. Pour aller plus loin sur ce point, il faut sans doute exploiter les informations fournies par l'enquête elle-même sur le comportement des individus en matière de fréquentation des centres. Le chiffre le plus demandé de l'enquête, dans le contexte français, sera sans doute une estimation de la taille de la population « sans-domicile », soit une estimation de la taille de $P(j)$. Au-delà des questions de comptage des liens déjà abondamment évoquées, cette estimation se heurte à plusieurs insuffisances de la base de sondage ainsi qu'à un caractère indirect de l'échantillonnage.

Le risque d'oublier certaines structures lors du dénombrement des centres est important. Même si l'inventaire est exhaustif, le décalage temporel entre cet inventaire et l'enquête à proprement parler rend probable l'apparition de nouvelles structures non recensées dans la base de sondage. Cela peut générer un biais dans la mesure où certains des individus qui fréquenteraient ces structures ne fréquenteraient par ailleurs aucun service de la base de sondage (On pourrait également s'attendre à un refus de coopération de la part des responsables de certains centres : pour l'enquête INSEE, le refus des institutions a été pratiquement inexistant (moins de 1 % de refus). Cela est dû en grande partie à un important travail de sensibilisation au moment du recensement des centres et juste avant l'enquête). Par ailleurs, l'absence de biais est conditionnée par un calcul correct des liens, les passages dans des centres non recensés ne devant pas être comptabilisés dans ces liens.

Les individus qui fréquenteraient des centres uniques (prestations) sont hors champ de l'enquête. (Leur celles où on se sera donné les moyens de compter les ment en dehors des heures « classiques ») (concrètement, Les individus qui fréquenteraient des centres uniques

— Au niveau des concepts, il demeure une difficulté puisque l'enquête doit se dérouler sur un mois et que la population-cible évolue au cours de la période.

— L'estimation de la taille de la population est donc partiellement fragile. Pour cette raison, on peut s'attendre à ce que les erreurs commises soient plus importantes pour les totaux que pour les moyennes.

6. ESTIMATION DANS LE CAS DE VARIABLES D'INTÉRÊT NON CONSTANTES AU COURS DE LA PÉRIODE D'ENQUÊTE

Certaines variables d'intérêt de l'enquête dépendent de la date d'observation, et ne sont donc pas constantes au cours de la période d'enquête. Ce peut être le cas de réponses à des questions portant sur la journée précédant l'interview, par exemple « Combien de repas avez-vous pris hier ? », « Combien de fois avez-vous dormi dans la rue la semaine dernière ? », etc. Les questions sur les liens sont également dans ce cas de figure. Il est donc important de voir dans quelle mesure on peut adapter le formalisme précédent à des estimations sur ce type de variables.

Si nous revenons à l'expression (3), il est facile de voir que la constance des y_k au cours de la période d'enquête est la condition qui permet de factoriser y_k et de faire apparaître les liens $r_k(j)$. On en déduit que le type de calcul mené ci-dessus est toujours valable pour des estimations portant sur des périodes plus courtes que la période sur laquelle les y_k sont constants.

Ainsi, pour des variables constantes sur un jour, on pourra parfaitement utiliser des estimateurs « un jour moyen ». Pour des variables constantes sur la semaine, on pourra utiliser des estimateurs « un jour moyen » ou « une semaine moyenne ».

7. CORRECTION DE LA NON-RÉPONSE TOTALE

Pour décrire complètement l'opération, il reste à préciser comment passer d'un jeu de probabilités d'inclusion (et donc de poids initiaux des prestations incluses dans

De manière identique, la moyenne de y sur la population des personnes qui fréquentent les services du champ de l'enquête un jour « moyen » est définie comme :

$$(6) \quad \psi = \frac{N}{\Theta} = \frac{\sum_{j=1}^J \Theta_j}{\sum_{j=1}^J N_j}$$

La définition des totaux ou moyennes une semaine donnée ou une « semaine moyenne » suit le même principe.

Pour estimer ces paramètres, il suffit d'adapter les formules de la section précédente, en constatant que les $r_k(j)$ doivent être remplacés par le nombre de prestations du champ de l'enquête dont la personne échantillonnée a bénéficié le jour (resp. la semaine) d'enquête.

Notons s_j l'échantillon des personnes interrogées le jour j , $r_k(j)$ le nombre de prestations de l'univers reçues par l'individu k le jour j uniquement, et $s_k(j)$ les prestations échantillonnées le jour j qui renvoient à l'individu k .

$$\Theta_j \text{ sera estimé par } \hat{\Theta}_j = \sum_{k \in s_j} y_k w_k$$

$$\text{où } \hat{w}_k = \frac{1}{\sum_{i \in s_j(j)} r_k(i)} w_i$$

Ici, les poids des individus dépendent du jour j . (Mais

pas les poids des prestations, w_i , qui sont fixés une fois pour toutes (en l'absence de non-réponse, il s'agit de l'inverse des probabilités de sélection des prestations)). Pour se convaincre de la différence entre Θ et X_j , l'analogue suivante est commode : on considère un guichet où chaque personne qui arrive doit remplir un dossier. Le cas de X_j correspond à un fonctionnement où une personne remplit un dossier la première fois où elle se présente au guichet, et n'en remplit plus les fois suivantes; le cas du « jour moyen » correspond à un fonctionnement où toute personne se présentant doit remplir un dossier, qu'elle soit déjà venue un jour précédent ou pas. Au bout d'une semaine par exemple, l'analyse des caractéristiques des personnes ayant rempli des dossiers sera très différente dans les deux cas : dans le deuxième cas, les personnes qui viennent souvent au guichet seront surreprésentées par rapport au premier cas. Il est possible de formaliser cette approche. Nous renvoyons le lecteur intéressé à Ardilly et Le Blanc (1999).

5.2 Estimation pratique des liens avec la base de sondage

Même si l'on se restreint à estimer des quantités de type « semaine moyenne » ou « jour moyen », il n'est pas en général possible de connaître les liens avec la base de sondage un jour donné (et a fortiori une semaine donnée ou sur toute la période de l'enquête).

5.3 Estimation sur l'ensemble de la période d'enquête

Estimer des totaux et des moyennes portant sur la population $P(j)$ peut apparaître comme un des objectifs de l'enquête. Cette estimation fait intervenir les liens des individus échantillonnés avec les prestations du champ de l'enquête. Il est donc nécessaire de modéliser l'évolution des liens au-delà d'une semaine, ou, ce qui revient au même, de modéliser le comportement de passage des individus dans les centres.

5.2.2 Estimation « une semaine moyenne »

Pour partager les poids, on garde tous les liens relatifs à la semaine. La première option décrite en 5.2.1 est évidemment à proscrire. Pour les estimations une semaine donnée, on peut prendre comme approximation pour les services fréquentés un jour j postérieur à la date d'entrée en services fréquentés par l'individu le jour $(j - 7)$. Cela est cohérent si l'on suppose qu'il existe une certaine saisonnalité des services fréquentés selon le jour de la semaine. Cela revient à remplacer dans les estimateurs la semaine civile de référence par une semaine glissante, c'est-à-dire les sept derniers jours à compter de la date d'interview. C'est l'option qui a été prise pour l'enquête, le questionnaire étant prévu pour récolter les liens sur les 7 jours précédant l'entrevue.

5.2.1 Estimation « un jour moyen »

5. PROBLÈMES D'ESTIMATION ET SOLUTIONS PRATIQUES DANS LA CAS D'UNE VARIABLE CONSTANTE

Dans les formules présentées précédemment, la connaissance des liens des personnes avec l'univers des prestations est indispensable. Or, ces quantités ne sont pas connues, pour plusieurs raisons :

- une raison théorique : parce que la collecte est étalée dans le temps, et qu'un individu interrogé en début de période ne peut pas prévoir les services qu'il va fréquenter après la date d'entretien (Notons que la collecte doit nécessairement être étalée dans le temps, si l'on vise une bonne couverture la population-cible; une collecte synchrone, même si elle était technique- ment réalisable, n'atteindrait pas l'ensemble de la population-cible mais seulement les personnes qui fréquentent les services à cette date),
- des raisons pratiques : parce que la mémoire des personnes interrogées fait défaut au-delà de quelques jours, et parce que la détection par l'enquêteur ou le concepteur d'enquête de prestations servies dans des centres n'appartenant pas à la base de sondage s'avère très difficile.

En pratique, il est donc impossible d'estimer sans biais un total d'intérêt sur la période de l'enquête (un mois) sans faire des hypothèses *a priori* (voir la section 5.3).

5.1 Estimation « un jour moyen », « une semaine moyenne »

On est donc amené à s'intéresser à des quantités qui font intervenir les liens sur une période courte, par exemple le jour ou la semaine. La population des personnes qui fréquentent les services du champ de l'enquête un jour f donné est $P_f = \cup_{c,j,f} c_{j,f}$. Introduisons les quantités suivantes relatives au jour f :

$$\Theta_f = \sum_{k \in P_f} y_k$$
$$N_f = \sum_{k \in P_f} 1 = \text{card}(P_f)$$

Si t = card (J) est le nombre de jours de la période de référence de l'enquête, nous définissons les paramètres d'intérêt suivants :

- le total de y sur la population des personnes qui fréquentent les services du champ de l'enquête un jour « moyen », dans le sens suivant :

$$\Theta = \frac{1}{J} \sum_{f=1}^J \Theta_f$$

(5)

Un cas particulier est le nombre de personnes qui fréquentent les services du champ de l'enquête un jour « moyen », $N = 1/J \sum_{f=1}^J N_f$.

Le calcul demeure complexe puisqu'il s'agit d'un plan à plusieurs degrés à probabilités inégales. Il est évidemment indispensable, pour ne pas sous-estimer la vraie variance, de conserver toutes les prestations dans les cas où plusieurs prestations échantillonnées renvoient au même individu.

4.4 Comparaison avec d'autres méthodes d'estimation

Ayant introduit des estimateurs de type « partage des poids », il convient de s'interroger sur une méthode d'estimation alternative, où l'on tenterait d'estimer directement les probabilités de sélection des individus dans l'échantillon. (L'estimateur du partage des poids n'est pas l'estimateur classique de Horvitz-Thompson : en effet, les poids de cet estimateur dépendent clairement de l'échantillon complet de prestations (voir la formule 4). Cette méthode peut en effet sembler plus naturelle. Deux remarques s'imposent :

- pour obtenir les probabilités de sélection des individus physiques, il n'est pas raisonnablement possible de procéder autrement que par l'intermédiaire des prestations qui renvoient à l'individu, à partir d'informations fournies par ce dernier quant à sa fréquentation des différents centres. Avec les notations précédentes, nous avons :
- $$\text{Prob}(k \in s_p) = \text{Prob}\left(\bigcup_{i \in \Pi(J); K(i) = k} i\right).$$
- La formule de Poïncaré permet d'exprimer cette probabilité à partir des probabilités d'inclusions simples, doubles, triples, etc, des prestations. Mises à part les probabilités d'inclusion simples, ces probabilités sont complexes car issues de tirages à probabilités inégales et sans remise. Il ne faut donc pas espérer obtenir une expression calculable pour $\text{Prob}(k \in s_p)$. Par contraste, la méthode du partage des poids est très simple à mettre en oeuvre.

- de manière plus structurelle, un problème vient du fait que les probabilités de sélection des prestations non échantillonnées ne sont pas connues *a priori*, du fait du tirage à plusieurs degrés. Aux degrés inférieurs, les probabilités de tirage dépendent du tirage précédent. Dans notre cas, on ne connaît pas la fréquentation des centres qui ne sont pas enquêtés. Pour obtenir la probabilité de sélection d'un individu, il faut connaître les probabilités d'inclusion de toutes les prestations dont il a bénéficié. Par contraste, une des forces de la méthode du partage des poids est de ne faire dépendre les poids des unités atteintes indirectement (ici les individus) que des poids des unités échantillonnées directement (les prestations). Cet avantage de la méthode est mentionné dans Lavalée (1995).

Cela suppose bien sûr que les informations données par la même personne enquêtée à deux endroits/jours différents soient les mêmes, ce qui est loin d'être acquis. En revanche, le repérage des individus peut s'avérer important pour limiter la non-réponse (voir section 7).

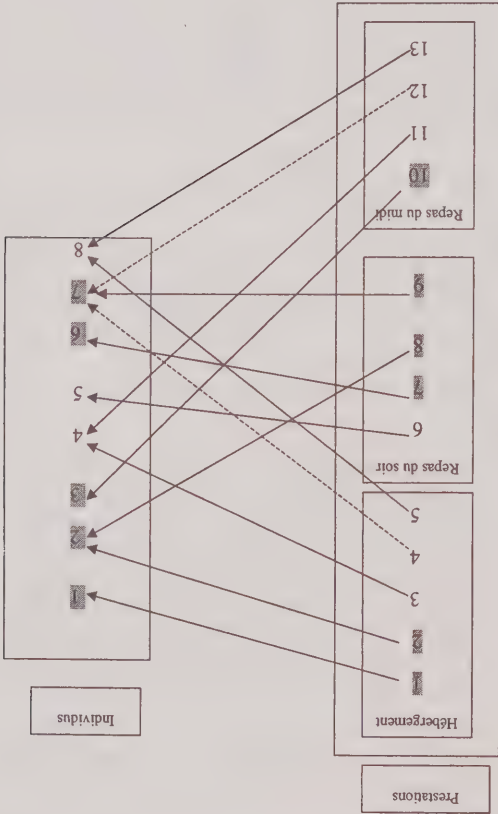


Figure 1. Les flèches représentent les liens entre les prestations et les individus. Les prestations dont l'identifiant est en gris sont échantillonnées. Elles renvoient aux individus en gris. Les traits en pointillé représentent les liens déclarés par l'individu 7 qui n'ont pas servi à l'échantillonnage.

4.2 Estimation d'un ratio

On suppose maintenant que l'on s'intéresse à l'estimation de la moyenne \bar{Y}_j (voir la formule 2). \bar{Y}_j peut être estimé par l'estimateur de Hajek,

$$\hat{\bar{Y}}_j = \frac{\bar{Y}_j}{\bar{N}_j}$$

$$\text{ou } \hat{N}_j = \sum_{k \in s_p} w_k$$

4.3 Calcul de variance

La variance des estimateurs présentés ci-dessus se calcule classiquement à condition de raisonner à partir des

est sans biais, où l'on a posé pour tout $k \in s_p$:

$$\bar{w}_k = \frac{1}{\sum_{s \in s_p} r_k(s)} \quad (4)$$

La formule (4) énonce simplement que le poids d'un individu est égal à la somme des poids des prestations qui ont servi à l'attraper », divisée par le nombre de liens avec la base de sondage, $r_k(s)$. On peut donc travailler directement sur les individus échantillonnés : pour chaque individu k , on calcule le poids \bar{w}_k , et on estime le total \bar{Y}_j par \bar{Y}_j .

La figure 1 donne un exemple fictif d'échantillonnage. L'univers des prestations contient 13 prestations, atteignant 8 personnes. 6 prestations sont échantillonnées. L'échantillon d'individus contient 5 personnes, l'individu numéro 2 ayant été « attrapé » par deux prestations différentes. Selon la formule (4), les poids des individus échantillonnés seront égaux à :

$$w_1 = w_2, w_3 = \frac{1}{2}(w_4 + w_8), w_5 = w_{10}, w_6 = w_7, w_7 = \frac{1}{3}w_9.$$

Si les prestations ont toutes le même poids égal à 1/36 (par exemple si les prestations ont été tirées par sondage aléatoire simple), le nombre de personnes ayant fréquenté les services pendant la durée de l'enquête est estimé par

$$\hat{Y}_j = \sum_{s \in s_p} w_k = \frac{6}{13} \left[1 + \frac{1}{2} \cdot 2 + 1 + 1 + \frac{3}{1} \right] = \frac{18}{169} \approx 9,39.$$

Dans le cas présent où la variable étudiée ne varie pas au cours de la période d'enquête, il est indifférent pour le biais de l'estimateur d'identifier les personnes fréquentant les prestations. Considérons en effet un individu « attrapé » par deux prestations différentes de poids w_1 et w_2 . Deux cas peuvent se produire en pratique :

— on repère que l'individu est le même; la pondération associée à cet individu sera égale à $(w_1 + w_2)/r_k(s)$, et le terme correspondant à l'individu dans l'estimateur sera égal à $y_k (w_1 + w_2)/r_k(s)$.

— on ne repère pas que l'individu a déjà été interrogé; on comptera deux individus différents; les pondérations associées à ces individus seront égales à $w_1/r_k(s)$ et $w_2/r_k(s)$, et le terme correspondant à ces deux pseudo-individus dans l'estimateur sera encore égal à $y_k (w_1 + w_2)/r_k(s)$.

Par exemple, y peut être la nationalité de l'individu, l'âge auquel il a terminé ses études, ou le nombre de centres qu'il a fréquentés le jour de l'entretien.

Nous serons par la suite amenés à distinguer deux types de variables :

- les variables fixes au cours de la période de référence de l'enquête (par exemple, l'âge de fin d'études).
- les variables qui varient au cours de la période de référence de l'enquête ($y_k = y_k(j)$). Le nombre de centres fréquentés le jour de l'enquête appartient à cette catégorie.

Nous traitons d'abord le cas des variables fixes au cours de la période de référence de l'enquête. La section 6 aborde brièvement le cas des variables qui varient au cours du temps.

4. ESTIMATION D'UN TOTAL OU D'UN RATIO DANS LE CAS OÙ LA VARIABLE D'INTÉRÊT EST CONSTANTE SUR LA PÉRIODE D'ENQUÊTE

Pour la commodité de l'exposé, nous ne faisons pas apparaître explicitement tous les degrés de tirage. Nous nous plaçons au niveau d'une agglomération échantillonnée au premier degré du tirage.

On note :

C : ensemble des centres de l'agglomération ouverts au moins un jour de la période d'enquête, repérés par l'indice c

$\Pi_{c,j,t}$: ensemble des prestations servies dans le centre c le jour j pendant l'intervalle de temps t , repérées par l'indice k

$\Pi_{j,t}$: ensemble des prestations servies dans l'intervalle de temps t , repérées par l'indice k

$P_{c,j,t}$: ensemble des personnes se présentant dans le centre c le jour j pendant l'intervalle de temps t , repérées par l'indice k

$P_{j,t}$: ensemble des personnes se présentant dans un des centres de l'agglomération le jour j pendant l'intervalle de temps t .

De la définition des intervalles de temps, il ressort qu'à chaque individu $k \in P_{j,t}$, correspond une et une seule prestation i . Ainsi, il existe une correspondance biunivoque entre $P_{j,t}$ et $\Pi_{j,t}$. Dit autrement, pour tout couple (j, t) , les $P_{c,j,t}$ sont disjointes. En revanche, $P_{c,j,t}$ et $P_{c^*,j,t}$ peuvent avoir une intersection non vide, dès que $t \neq t^*$.

La population d'intérêt s'écrit alors

$$P(J) = \bigcup_{c \in C} P_{c,j,t} = \bigcup_{c \in C} \left(\prod_{j,t} P_{c,j,t} \right).$$

Le point central du raisonnement consiste à exprimer le total d'une variable sur la population des *individus* (qui est

notre total d'intérêt) comme le total d'une autre variable sur la population des *prestations* (qui sont les unités échantillonées), l'estimation de ce dernier ne posant aucune difficulté particulière. Pour obtenir ce résultat, on peut recourir à un raisonnement direct, ou appliquer la méthode du partage des poids, l'un ou l'autre pouvant sembler plus naturel.

En raisonnant directement, nous définissons l'application K , qui à toute prestation i servie durant la période de référence J dans l'ensemble des centres du champ de l'enquête, associe l'individu bénéficiaire de cette prestation.

L'ensemble des prestations servies durant la période de référence dans l'ensemble des centres du champ de l'enquête. Pour tout $k \in P(J)$, on définit $r_k(J) = \text{card}(K^{-1}(k))$, le nombre de prestations servies à l'individu k durant la période J dans l'ensemble des centres du champ de l'enquête, que nous appellerons aussi « nombre de liens ».

On a l'égalité fondamentale :

$$Y_J = \sum_{k \in P(J)} y_k = \sum_{k \in P(J)} \frac{y_{K(i)}}{r_{K(i)}(J)}.$$

En effet, la variable y prenant la même valeur pour toutes les prestations i « pointant » sur l'individu k , c'est-à-dire telles que $K(i) = k$, le membre de droite peut s'écrire

$$\sum_{k \in P(J)} \left[\sum_{i \in \Pi(J); K(i) = k} \frac{y_k}{r_k(J)} \right] = \sum_{k \in P(J)} \frac{y_k}{r_k(J)} \left[\sum_{i \in \Pi(J); K(i) = k} 1 \right].$$

Mais la quantité entre crochets est le nombre de prestations servies à l'individu k durant la période J , soit $r_k(J)$, ce qui prouve l'égalité.

On peut alors voir $r_{K(i)}(J)$ comme attaché à la prestation i correspondant et noter y_i au lieu de $y_{K(i)}$ et $r_i(J)$ au lieu de $r_{K(i)}(J)$. En notant $z_i = y_i/r_i(J)$, $Z = \sum_{i \in \Pi(J)} z_i$, on a

$$Z = \sum_{i \in \Pi(J)} z_i = \sum_{i \in \Pi(J)} \frac{y_i}{r_i(J)}.$$

La formule (3) n'est autre que celle du partage des poids. Le raisonnement ci-dessus est d'ailleurs celui qui sous-tend cette méthode. (Seules les notations changeant, la méthode du partage des poids décrit les liens entre la population échantillonnée et la population d'intérêt par une matrice d'échantillonnage pouvant « pointer » vers plusieurs unités de la population d'intérêt). Le principe de cette dernière est rappelé en annexe 1.

4.1 Estimation d'un total

Supposons maintenant que l'on dispose d'un échantillon π de prestations, auquel est associé un jeu de poids $(w_i)_{i \in \pi}$. Nous supposons ces poids sans biais (il s'agit de l'inverse des probabilités d'inclusion des prestations dans

le même centre. Il se trouve par ailleurs que la largeur d'un intervalle assurant une telle propriété correspond à la durée au cours de laquelle on peut raisonnablement demander à un enquêteur d'interroger sur place (soit 2 à 3 heures au maximum). (On remarquera que les accueils de jour ne font pas partie des services retenus dans le champ de l'enquête. Cette restriction de champ correspond à deux préoccupations. D'une part, il serait très difficile de découper la journée en intervalles de temps de trois ou quatre heures et de collecter les liens sur la base de ce découpage (l'effort de mémoire demandé à l'enquêteur serait considérable et n'a pas paru raisonnable aux concepteurs de l'enquête). D'autre part, les fréquentations de ces services sont très peu prévisibles. On a voulu éviter à une équipe d'enquêteurs de se déplacer et de ne réaliser aucun entretien faute de fréquentation.)

En fait, il n'y a pas de différence fondamentale entre l'échantillonnage des centres et l'échantillonnage des périodes de temps : les unités pertinentes à considérer sont les triples (c, j, t) correspondant au croisement d'un centre, d'un jour et d'un intervalle de temps. Certaines cases du tableau croisant « temps » et « centres » seront éliminées *a priori* avant le tirage, soit parce que le centre est fermé durant le créneau horaire considéré, soit parce que la fréquentation y est manifestement très faible. (Dans ce dernier cas, il faut prendre garde à l'éventuelle restriction du champ couvert, s'il s'agit que des personnes ne fréquentent que ce centre et ne sont présentes que dans ce créneau horaire. Si ces dernières sont atypiques, des biais seront introduits dans les estimations.)

Le mode de tirage retenu a consisté à tirer au hasard des triples (centres, jour, intervalles de temps) proportionnellement à la taille des centres obtenue lors du recensement des centres. (En pratique, des regroupements d'intervalles de temps ont eu lieu dès lors qu'un centre était échantillonné plus de quatre fois au cours de la période d'enquête, pour des raisons d'acceptabilité par les responsables de centre.) Une stratification par type de centre a été effectuée. (Pour les services d'hébergement, une stratification sur le critère hommes/seulement/femmes/seulement/accueil mixte a été introduite.) Toutefois, cette stratification « de précaution », ne portant pas directement sur les unités d'observation, n'aura été utile que si le comportement des personnes diffère sensiblement selon le type de centre où on les trouve.

2.3.3 Tirage des prestations

Ce dernier degré consiste à achever l'échantillonnage des prestations, c'est-à-dire à tirer des personnes dans un centre sélectionné un jour donné dans un intervalle de temps donné. Les données recueillies lors du recensement de ces centres ne suffisent pas en général à constituer une base de sondage de prestations. Dans certains centres d'hébergement, il peut exister des listes ; c'est le cas le plus favorable, un tirage des personnes pouvant être conduit à partir de ces listes. En revanche, dans la majorité des centres (par

exemple, dans un point-soupe), on ne connaît même pas le nombre de personnes qui vont se présenter durant un intervalle de temps donné : on ne peut donc pas faire de base de sondage des prestations. L'échantillonnage des prestations s'effectue à probabilités égales. Comme traditionnellement dans les sondages à plusieurs degrés, tirer un nombre constant de prestations (dernier degré) permet d'assurer des probabilités de tirage constantes, et donc de limiter les risques d'explosion de variance. En pratique, la méthode de tirage retenue varie d'un type de centre à un autre, selon la topographie des lieux : liste existante, file d'attente, arrivées espacées dans le temps, population « groupée » sans ordre dans un même lieu au même moment, etc. Elle tient aussi compte du nombre maximal d'interviews raisonnablement assurables par le ou les enquêteurs pendant l'intervalle de temps de l'enquête, et du fait qu'il n'est pas souhaitable de retenir des personnes échantillonnées trop longtemps après la fermeture d'un centre ou l'arrêt de distribution de repas, sous peine d'augmenter la non-réponse. Dans tous les cas, un « dénombreur » compte pendant la période d'échantillonnage le nombre N de prestations services. Ce rôle est essentiel pour déterminer la probabilité de tirage des prestations échantillonnées. Parallèlement, il procède à un tirage de type systématique (Dans l'idéal, le tirage devrait être assuré par une autre personne (ou « échantillonneur »), afin d'éviter les erreurs de mesure sur la fréquentation. Des raisons budgétaires ont conduit à ne pas retenir cette solution.) selon la méthode suivante :

- dans les centres où une liste est disponible, on tire n prestations, n étant fixé avant l'enquête.
- dans les centres où aucune liste n'est disponible, on tire les prestations avec un taux de sondage f fixe. f est déterminé selon le nombre de prestations attendues N et le nombre de prestations que l'on désire échantillonner n , afin d'assurer des probabilités de tirage égales. Dans ce cas, la taille de l'échantillon est inconnue *a priori*.

3. LES PARAMÈTRES D'INTÉRÊT

Les quantités d'intérêt sont essentiellement du type totaux ou ratios. Nous désirons estimer un total relatif à une variable y définie sur la population $P(j)$.

$$Y_j = \sum_{k \in P(j)} y_k. \quad (1)$$

Un cas particulier de ces totaux est l'effectif de $P(j)$, $N_j = \text{card}(P(j)) = \sum_{k \in P(j)} 1$. Nous désirons également estimer la moyenne de y dans la population de référence,

$$\bar{Y}_j = \frac{N_j}{Y_j} = \frac{1}{N_j} \sum_{k \in P(j)} y_k. \quad (2)$$

2.2 La population de référence

Le plan de sondage de l'enquête comprend trois degrés : tirage d'agglomérations, tirage de centres et d'intervalles de temps, et enfin tirage de prestations.

2.3.1 Tirage des agglomérations

Le premier degré du plan de sondage consiste à tirer des agglomérations, proportionnellement à un critère de taille défini comme une combinaison de la population des agglomérations et des capacités d'accueil telles qu'elles ont pu être recensées dans les fichiers des associations et les fichiers du Ministère de la Santé. Ce premier degré de tirage est effectué plusieurs mois avant les autres. Ce décalage s'impose car le recensement exhaustif des centres et des informations les concernant (type de service rendu, capacité moyenne, jours d'ouverture, ...) est entrepris sur les agglomérations tirées. Cette opération est réalisée en deux fois : une enquête lourde l'année précédant la collecte, et une mise à jour juste avant le début de la collecte. On obtient ainsi une base de sondage de centres. Cette base joue un rôle fondamental : des personnes qui ne fréquenteraient que des centres non recensés seraient de fait non échantillonnables.

2.3.2 Tirage des centres, des jours et des intervalles

Pour des raisons pratiques, il n'est pas possible d'enquêter l'ensemble des centres et de maintenir sur le terrain, dans un centre donné, un enquêteur durant une journée entière. Enfin, on ne peut interroger toutes les personnes dans un centre. Il est donc incontournable d'échantillonner :

- des centres dans les agglomérations tirées (indice c)
- des jours d'enquête pendant la période de collecte (indice j)
- des intervalles de temps pendant les jours d'enquête (indice i).
- des personnes au sein d'un (centre, jour, intervalle de temps) tiré.

Pour des raisons théoriques, les intervalles de temps sont définis de façon qu'un individu ne puisse pas bénéficier de deux prestations différentes durant cet intervalle de temps (par exemple, un de ces intervalles de temps est la période de 11 à 14 heures). En effet, la mesure des liens avec la base de sondage ne peut raisonnablement s'effectuer qu'en permettant aux personnes interrogées de repérer facilement dans le temps et l'espace les prestations qui leur ont été servies au cours de la période d'enquête. Pour les centres offrant des repas, un intervalle de temps recouvrira les repas du midi et un intervalle les repas du soir. On considère qu'une personne ne peut fréquenter qu'un seul centre durant l'intervalle de temps correspondant au repas de midi, faute de quoi il faudrait lui demander si elle n'a pas déjà pris un repas ailleurs, ou si elle ne mange pas deux fois dans

La question de la détermination de J revient finalement à savoir si on s'intéresse plutôt à une notion de sans-domicile « à un instant donné » (J plutôt court), ou à une notion de sans-domicile retenue par l'INSEE constitue un compromis entre les deux.

La population de référence

La caractéristique principale des services considérés est qu'ils sont fournis dans un lieu précis ; ce lieu est appelé par la suite *centre*. À un centre donné correspond un ou plusieurs types de services. L'unité statistique échantillonnée, que nous appellerons par la suite *prestation*, sera définie comme un quadruplet (service, jour, intervalle de temps, personne) : il s'agit d'un service de type donné dans un centre donné, un jour donné, dans un intervalle de temps donné, à une personne donnée. Une personne peut bien sûr bénéficier de plusieurs prestations la même journée, et *a fortiori* une semaine donnée ou pendant le mois d'enquête.

La période de référence de l'enquête s'étend sur un mois (du 15 janvier au 15 février 2001). On note J l'ensemble des jours de la période de référence de l'enquête, repérés par l'indice j .

Le champ géographique de l'enquête est celui des agglomérations de plus de 20 000 habitants.

Les prestations dans le champ de l'enquête sont celles qui reviennent d'un des deux types de services retenus : repas et hébergement, dès lors qu'ils sont assurés au moins une journée pendant la période de référence de l'enquête.

La population de référence, notée $P(j)$, est constituée des personnes qui ont bénéficié d'au moins une prestation du champ de l'enquête pendant la période de référence.

Cette population d'intérêt dépend fondamentalement de la période de référence. Sa taille croît avec la longueur de cette période, mais « moins vite » que le temps : en effet, d'un jour sur l'autre, on retrouve certaines personnes dans les centres. En réalité, l'évolution de $P(j)$ avec j est complexe, car deux phénomènes distincts interviennent, dont on peut penser qu'ils ont des temps caractéristiques différents :

- la population « sans-domicile » à un moment donné ne fréquente qu'épisodiquement les centres de la base : période de temps où toutes les personnes de cette population ont au moins une fois recours à des services, (cette période n'est pas connue, mais il est admis en France, « à dire d'expert », que la population non couverte pendant un mois d'hiver complet est de taille négligeable).
- la population « sans-domicile » se renouvelle dans le temps. D'une année sur l'autre, des entrées et des sorties, sans doute nombreuses, interviennent, liées aux mouvements démographiques ou aux évolutions conjoncturelles ou structurelles de la société (entrées et sorties de situations de précarité).

La question de la détermination de J revient finalement à savoir si on s'intéresse plutôt à une notion de sans-domicile « à un instant donné » (J plutôt court), ou à une notion de sans-domicile retenue par l'INSEE constitue un compromis entre les deux.

Echantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français

PASCAL ARDILLY et DAVID LE BLANC¹

RÉSUMÉ

L'INSEE a réalisé en 2001 une enquête destinée à mieux connaître la population sans domicile. En l'absence de base de sondage permettant d'atteindre directement les personnes sans domicile, le principe de l'enquête est d'échantillonner des prestations qui leur sont destinées et d'interroger les individus qui bénéficient de ces prestations. Lorsque l'on désire pondérer les observations individuelles issues de l'enquête, une difficulté surgit du fait qu'un individu peut bénéficier de plusieurs prestations pendant la période de référence considérée. Cet article montre comment il est possible d'appliquer la méthode du partage des poids pour résoudre ce problème. Dans ce type d'enquête, une même variable peut donner lieu à plusieurs paramètres d'intérêt, correspondant à des populations variant avec le temps. À chaque définition des paramètres correspond un jeu de poids. L'article insiste particulièrement sur le calcul de poids « un jour moyen » et « une semaine moyenne ». On donne également des éléments sur les données de fréquentation à collecter et la correction de la non-réponse.

MOTS CLÉS : Partage des poids; base incomplète; personnes sans-domicile.

1. INTRODUCTION

L'INSEE a réalisé en 2001 une enquête destinée à mieux

connaître la population des sans-domicile. Cette enquête est la première enquête représentative de ce type en France (Une enquête de ce type a été menée aux États-Unis en 1991 par le *Research Triangle Institute* (RTI) dans la région métropolitaine de Washington (RTI 1993)). Le principe de l'enquête est d'atteindre les personnes sans domicile par le biais de prestations qui leur sont destinées, hébergement de nuit et repas. Évidemment, une personne peut fréquenter une ou plusieurs prestations de la base de sondage pendant la période de référence considérée, ce qui pose une difficulté lorsque l'on désire pondérer le fichier d'observations individuelles issu de l'enquête. Dans cet article, nous montrons comment la méthode du partage des poids peut être appliquée à ce problème. Dans ce type d'enquête, contrairement à la plupart des enquêtes traditionnelles auprès des ménages, une même variable peut donner lieu à plusieurs paramètres d'intérêt, correspondant à différents concepts de population : les plus utilisés par les praticiens sont les paramètres « un jour moyen » et « une semaine moyenne ». À chaque définition des paramètres correspond un jeu de poids. Nous définissons précisément ces concepts, et insistons particulièrement sur le calcul pratique des poids correspondants. Le plan de l'article est le suivant : nous rappellerons d'abord les objectifs de l'enquête, sa population de référence et son plan de sondage. Nous introduisons ensuite les paramètres d'intérêt et dérivons les estimateurs de ces paramètres issus de la méthode du partage des poids. Nous décrivons la mise en oeuvre pratique de calculs de poids « un jour moyen » et « une semaine moyenne ». Enfin, nous donnons des considérations pratiques sur la correction de la non-réponse.

2.1 Objectifs de l'enquête

2. L'ENQUÊTE « SANS DOMICILE »

L'enquête réalisée par l'INSEE en février 2001 vise à mieux connaître la population des « sans-domicile ». Cette population est généralement définie par défaut comme l'ensemble des personnes qui ne disposent pas d'un domicile fixe. Cette population échappe aux enquêtes traditionnelles auprès des ménages menées par l'Institut, celles-ci reposant sur une base de sondage de logements. En l'absence d'une base de sondage recensant cette population, le principe de l'enquête consiste à atteindre la population visée par le biais de prestations destinées aux personnes en difficulté, correspondant à des services d'hébergement et de repas. Ces prestations sont fournies sur des bases temporelles qui varient selon leur nature : les repas sont fournis chaque jour midi et soir, les nuitées une fois par jour. Cet échantillonnage indirect induit deux distorsions entre la population visée initialement et la population atteinte par l'enquête. D'une part, la population visée n'est pas atteinte dans sa totalité : seuls ses membres qui utilisent les prestations rentrant dans le champ de l'enquête sont potentiellement échantillonnables. D'autre part, la population atteinte par l'enquête contient des personnes qui n'appartiennent pas à la population visée initialement, dans la mesure où les services destinés en priorité aux personnes sans domicile sont aussi utilisés par des personnes qui vivent en ménage ordinaire mais sont en situation précaire (c'est surtout le cas pour les repas). Dans tout ce qui suit, tout en gardant cette distinction à l'esprit, on parlera cependant parfois de « sans-domicile » pour désigner les personnes fréquentant les prestations du champ de l'enquête.

¹ Pascal Ardilly et David Le Blanc, Institut National de la Statistique et des Études Économiques, 18 boulevard Adolphe Pinard, 75675, Paris, Cedex, France.
Courrier électronique : pascal.ardilly@insee.fr, leblanc@ensae.fr.

BIBLIOGRAPHIE

ANGOSS SOFTWARE (1995). Knowledge Seeker IV for Windows – User's Guide. ANGOS Software International Limited.

CHAPMAN, D.W., BAILEY, L. et KASPRZYK, D. (1986). Méthodes de compensation de la non-réponse au U.S. Bureau of the Census. *Techniques d'enquête*, 12, 167-187.

DEVILLE, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Actes du Congrès de la Société Statistique du Canada, Recueil de la Section des méthodes d'enquête*, 103-110.

DEVILLE, J.-C., et SARNADAL, C.-E. (1992). Calibration estimators and generalized raking techniques in survey sampling. *Journal of the American Statistical Association*, 87, 418, 376-382.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

LAVIGNE, M., et MICHAUD, S. (1998). Aspects généraux de l'Enquête sur la dynamique du travail et du revenu. Document de recherche de l'EDTR, Statistique Canada, catalogue 98-05.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de la Statistique*, 54, 137-139.

MICHAUD, S., MORIN, Y., CLERMONT, Y. et LAFLAMME, G. (1998). Issues in the design of a survey to measure child development : The Experience of the Canadian National Longitudinal Survey of Children and Youth. Statistique Canada, document interne.

PLATEK, R., SINGH, M.P. et TREMBLAY, V. (1978). Adjustment for nonresponse in surveys. *Survey Sampling and Measurement*. N.K. Namboodiri, Ed. Academic Press, 157-174.

RIZZO, L., KALTON, G. et BRICK, M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

SINGH, A. C., WU, S. et BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Actes de la Conférence de l'American Statistical Association*, 396-401.

REMERCIEMENTS

Les auteurs aimeraient remercier M. Hladky, M. Latouche, C. Nadeau et N. Tremblay pour leur importante contribution à ce projet.

En particulier, il semble que plus la valeur du terme R_{01} est élevée, plus la réduction du biais obtenue en utilisant les GRH est importante. Étant donné la difficulté d'obtenir une estimation fiable du biais de non-réponse dans une enquête, la relation identifiée entre la taille de la composante R_{01} et la réduction du biais suggère d'utiliser R_{01} comme un outil pour évaluer les méthodes d'ajustement de non-réponse. Pour ce faire, il faut tout d'abord déterminer R_{01} pour différents ensembles de GRH. Ensuite, l'ensemble affichant la plus grande valeur de R_{01} est sujet à être plus efficace que les autres alternatives pour réduire le biais de non-réponse, pour la plupart des variables d'intérêt.

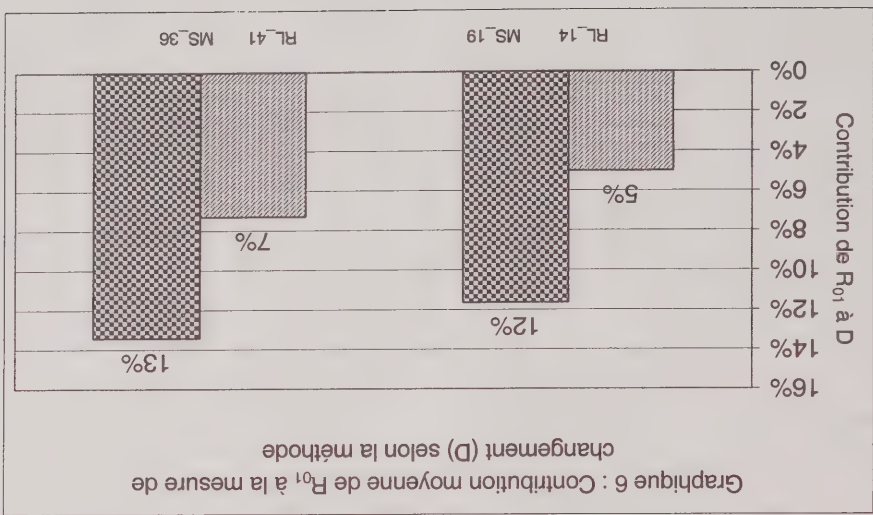
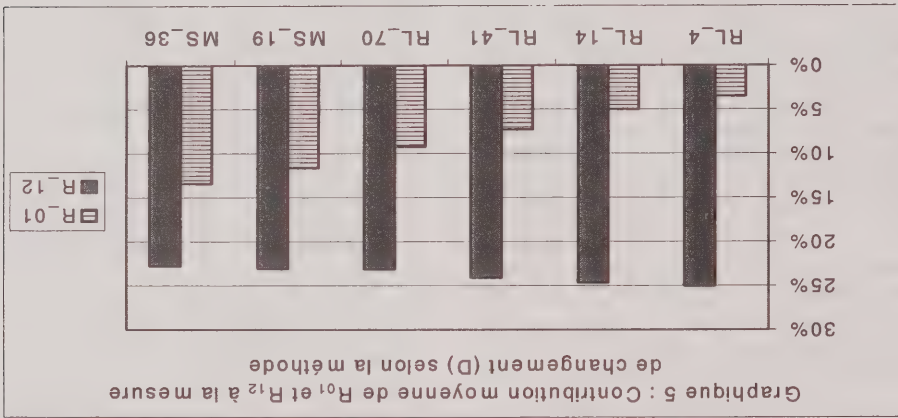
La mesure de changement présentée pourrait aussi être utilisée pour comparer différentes stratégies de calage. Dans ce cas, l'ajustement de non-réponse pourrait demeurer le même pour toutes les méthodes de poststratification à l'étude. Une étude détaillée du comportement du terme R_{12} pourrait être effectuée et permettrait sans doute de tirer certaines conclusions comme la présente étude a permis de tirer des conclusions à propos du terme R_{01} . Ce genre d'étude n'aurait pas nécessairement besoin de se limiter au contexte longitudinal mais pourrait fort bien s'effectuer dans le cas d'une enquête transversale. De même, la mesure de changement pourrait aussi s'avérer utile pour évaluer différentes méthodes d'ajustement de non-réponse d'enquêtes transversales.

Tout comme observé lors de l'étude empirique, le profil même que celui de la mesure elle-même. Cela montre que les variations de D dépendent directement de R_{01} . Le graphique 6 permet de comparer la RL et le MS en présentant la contribution moyenne de R_{01} à la mesure de changement pour les méthodes avec un nombre sensible-ment équivalent de GRH. Comme pour l'EDTR, les résultats montrent que la non-réponse semble être mieux ciblée avec la méthode du MS qu'avec la RL.

Contrairement à l'étude de simulation de l'EDTR, le biais n'a pas été évalué étant donné qu'aucune source externe de données n'était disponible pour fins d'évaluation.

7. CONCLUSION

Ce document met en relief le fait que le choix des GRH et de la méthode pour les définir dépendent : i) de la disponibilité de l'information auxiliaire, ii) du souci de réduire le biais de non-réponse pour toutes les estimations, et iii) du temps et des contraintes opérationnelles. L'étude empirique, de même que les données de l'ENLEJ, ont démontré que la méthode du MS semble meilleure que celle de la RL pour réduire le biais de non-réponse. Les résultats ont également démontré que la mesure de changement proposée peut être un outil très utile pour comparer différentes stratégies de pondération.



6. APPLICATION AUX DONNÉES DE

L'ENQUÊTE NATIONALE
LONGITUDINALE SUR LES
ENFANTS ET LES JEUNES (ENLEJ)

Dans cette section, la majorité des analyses réalisées à l'aide de la RL et du MS lors de l'étude empirique avec les données de l'EDTR sont reproduites en utilisant l'information obtenue de l'ENLEJ. Tout comme l'EDTR, l'ENLEJ est une enquête longitudinale auprès des ménages. Elle a débuté en 1994 et son but est de recueillir de l'information pour analyser les politiques et développer des programmes portant sur les facteurs critiques qui affectent le développement des enfants au Canada (voir Michaud, Morin, Clermont et Laflamme 1998).

6.1 Description et analyse des résultats de l'application

Pour cette étude, les méthodes utilisées sont les suivantes : RL_{-i}, où $i=4, 14, 41, 70$ avec respectivement $q=2, 4, 6, 8$ variables, et MS_{-i}, où $i=19, 36$ avec des seuils de signification de 0,001 et 0,005 respectivement. Les deux mêmes contraintes imposées pour l'EDTR ont été réappliquées lors de la création des GRH. Pour chacune des méthodes à l'étude, la même poststratification a été utilisée (22 groupes d'âge-sexe par province). Contrairement à l'étude empirique basée sur l'EDTR, seules les données recueillies aux deux premières vagues de l'ENLEJ ont été utilisées. Aucune simulation n'a été effectuée et les poids initiaux n'ont pas été normés ($\sum w_{0k} = N < N$). Il est à noter que la sous-couverture de l'ENLEJ est d'environ 13 % et la non-réponse aux alentours de 8 %.

Les résultats présentés au tableau 3 permettent de tirer des conclusions similaires à celles obtenues lors de la simulation (tableau 1). Cependant, on observe que la contribution relative de R_{01} à la mesure de changement est plus faible pour l'ENLEJ que dans le cas de l'EDTR. Ce résultat indique que l'ajustement de non-réponse de l'EDTR cause de plus grands changements individuels dans les poids, entraînant une plus grande contribution du terme R_{01} . Dans

Tableau 3
Valeur de D , de chaque composante et de leur contribution (en %) à la mesure de changement pour chacune des six méthodes d'ajustement de la non-réponse

Méthode	D	R_{01}	R_{01}/D	R_{12}	R_{12}/D	R_{int}	R_{int}/D	R_{int}/D	G	G/D
RL_4	0,1475	0,0052	3,51	0,0369	25,05	-4,63	-0,31	0,1058	71,76	64,52
RL_14	0,1497	0,0075	5,00	0,0367	24,69	-5,50	-0,37	0,1058	70,68	69,18
RL_41	0,1530	0,0112	7,29	0,0369	24,13	-9,16	-0,60	0,1058	67,67	65,81
RL_70	0,1564	0,0144	9,21	0,0362	23,13	-0,19	-0,01	0,1058	67,67	65,81
MS_19	0,1608	0,0187	11,63	0,0371	23,07	-8,24	-0,51	0,1058	65,81	64,52
MS_36	0,1640	0,0220	13,41	0,0373	22,76	-11,30	-0,69	0,1058	64,52	64,52

le cas de l'ENLEJ, l'ajustement de non-réponse n'a donc pas d'effet majeur sur les changements individuels dans les poids, contrairement à ce qui a été observé dans le cas de l'EDTR.

La contribution relative de R_{12} à la mesure de changement est plus élevée pour l'ENLEJ que pour l'EDTR. Ce résultat indique que la poststratification plus raffinée de l'ENLEJ entraîne de plus grands changements individuels dans les poids, ce qui se traduit par une plus grande contribution de R_{12} . L'ENLEJ profite donc beaucoup de la poststratification alors qu'elle a une importance beaucoup moindre pour l'EDTR.

Pour ce qui est de R_{int} , tout comme pour l'EDTR, sa contribution à la mesure de changement est négligeable. Contrairement à l'EDTR, le signe de R_{int} est négatif ce qui signifie que l'interaction entre R_{01} et R_{12} est négative. Pour ce qui est de G , comme dans le cas de l'EDTR, il est la principale source de contribution à la mesure de changement. Dans le cas de l'ENLEJ, G inclut non seulement le changement de poids moyen causé par l'ajustement de non-réponse mais aussi le changement de poids moyen causé par la correction pour la sous-couverture via la poststratification.

En comparant tous ces résultats, on s'aperçoit que les deux enquêtes se ressemblent beaucoup étant donné que $R_{int} \approx 0$ et que la somme des contributions à la mesure de changement de R_{01} et R_{12} se situe approximativement autour de 35 % dans chacun des cas. Par contre, l'ENLEJ est aussi très différente de l'EDTR car R_{12} prédomine pour la première alors que R_{01} prédomine pour la seconde.

Tout comme pour l'EDTR, D augmente avec le nombre de GRH et cette mesure est plus élevée pour le MS que pour la RL. La valeur de D est d'ailleurs plus élevée pour l'ENLEJ que pour l'EDTR principalement à cause de la sous-couverture de l'ENLEJ, ce qui a pour effet d'augmenter G et conséquemment D .

La contribution moyenne de R_{01} pour la RL et le MS augmente avec le nombre de GRH tandis que celle de R_{12} diminue (graphique 5). La contribution de R_{01} est également plus élevée pour le MS que pour la RL, contrairement à la contribution de R_{12} qui est plus petite pour le MS que pour la RL.

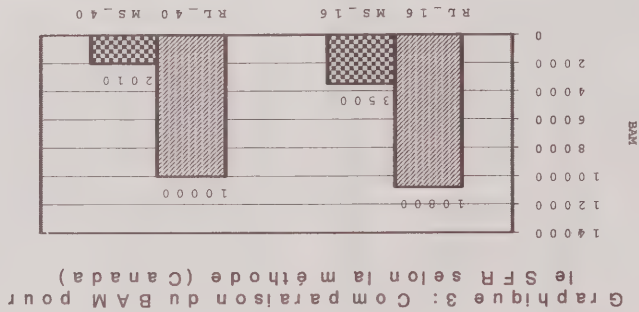
5.2.3 Estimations de variance

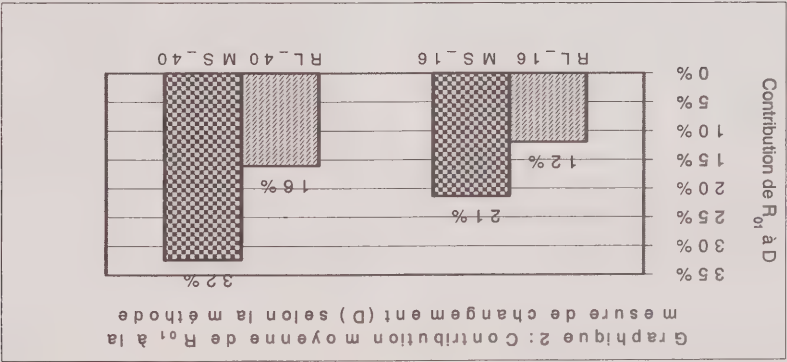
Des estimations de variance ont été produites pour les trois variables d'intérêt à l'aide de la méthode du jackknife. Pour le SFR (graphique 4), la variance moyenne des estimations est de façon approximative la même, peu importe la méthode étudiée. Il y a cependant une légère baisse

lorsque le nombre de GRH augmente, autant pour la RL que pour le MS. De plus, selon l'étude empirique, les estimations de la variance moyenne du MS sont légèrement plus petites que celles de la RL. On constate qu'une plus grande dispersion dans les poids (une valeur de D plus élevée) n'entraîne donc pas une augmentation de la variance.

Tableau 2
BRM (en %) pour différentes variables selon les méthodes étudiées – Canada

Variable	MÉTHODE UTILISÉE						
SFR	1-GRH	RL-4	RL-16	RL-40	MS-16	MS-25	MS-40
RT		0,37	0,43	0,37	0,31	0,14	0,12
ST		-0,32	-0,09	-0,06	-0,06	-0,006	-0,005
		-0,44	-0,13	-0,15	-0,19	-0,14	-0,10
							-0,09





Pour les 100 répétitions, des estimations à l'échelle nationale ont été produites pour les trois variables suivantes : « personne vivant ou non dans une famille dont le revenu est inférieur au seuil de faible revenu (SFR) », « revenu total de la personne (RT) » et « salaires et traitements de la personne (ST) ». Le BRM de chaque estimation a été calculé pour les huit méthodes à l'étude. Étant donné la grande taille de l'échantillon, le faible taux de non-réponse (10 %) et le fait qu'un grand nombre de totaux de contrôle ont été utilisés pour la poststratification, le BRM est très faible (voir tableau 2) pour chacune des méthodes utilisées.

On observe au tableau 2 que pour les trois variables, le BRM est à peu près constant pour le MS peu importe le nombre de GRH utilisés. De même, pour la RL, le BRM du RT et des ST est à peu près constant peu importe le nombre de GRH utilisés. Par contre, pour le SFR, le BRM de la méthode RL_4 est nettement plus petit que le BRM des trois autres méthodes de la RL. Ceci peut s'expliquer par le fait que le SFR est une variable dérivée à partir de plusieurs autres variables, contrairement au RT et aux ST qui sont des variables observées. Le BRM des trois variables pour la méthode 1_GRH est beaucoup plus grand que le BRM produit par le MS et par la RL, sauf pour le SFR où dans ce

cas, le BRM est environ équivalent au BRM de la RL. Il semble donc que la méthode 1_GRH performe moins bien que le MS et la RL. Dans le meilleur des cas, elle est environ équivalente à la RL. Contrairement au MS, on note que la progression du BRM n'est pas strictement décroissante pour la RL à mesure que le nombre de GRH augmente.

Malgré le fait que le BRM soit minime pour les variables étudiées pour le Canada, il peut augmenter rapidement pour de petits domaines. Pour la présente étude, d'autres domaines ont également été étudiés. Quoique certains écarts soient observés pour plusieurs d'entre eux, il semble que le BRM du MS soit généralement plus petit que le BRM de la RL et la méthode 1_GRH. Une étude plus approfondie sur un plus grand nombre de variables d'intérêt et de domaines serait bénéfique afin de corroborer ces conclusions.

Tel que mentionné précédemment, les changements individuels dans les poids causés par l'ajustement de non-réponse sont supérieurs pour le MS que pour la RL (voir graphique 2). Ceci semble indiquer que le MS est plus efficace pour réduire le BA de non-réponse, pour un nombre fixe de GRH. Le graphique 3 confirme cette observation; on y observe que le BAM pour le SFR est inférieur selon la méthode du MS que pour la RL.

où α est une constante, et par conséquent, $r_{01k} = 1$ pour tout $k \in r$ et $R_{01} = 0$. On remarque également que D augmente à mesure que le nombre de GRH augmente peu importe la méthode, RL ou MS. Ainsi, plus il y a de GRH pour pallier la non-réponse, plus le changement total que subissent les poids est grand. De plus, les valeurs de D sont plus élevées pour le MS que pour la RL.

Pour la RL et le MS, la contribution de R_{01} à la mesure de changement est de plus en plus grande à mesure que le nombre de GRH augmente étant donné qu'un plus grand nombre de GRH cible mieux la non-réponse. Par conséquent, l'ajustement de non-réponse prend de plus en plus d'importance et par le fait même, les poids varient de plus en plus. De plus, la contribution de R_{01} à la mesure de changement est beaucoup plus importante pour le MS que pour la RL. Ceci indique que le MS semble mieux modéliser la non-réponse et isoler mieux les tendances particulières que la RL.

Quant à R_{12} , il est presque constant, peu importe la méthode et le nombre de GRH utilisés. Cependant, malgré le fait qu'il varie très peu, sa contribution à la mesure de changement diminue à mesure que le nombre de GRH augmente. Ceci s'explique par le fait qu'une plus grande variation des poids survient lors de l'ajustement de non-réponse et que les modifications que la poststratification apporte aux poids sont de moins en moins importantes à mesure que le nombre de GRH augmente.

Dans le cas de R_{int} , sa valeur est négligeable et sa contribution à la mesure de changement est très faible. Ceci signifie que l'interaction entre l'ajustement de non-réponse et la poststratification est quasiment nulle. Enfin, G demeure constant, peu importe la méthode et le nombre de GRH utilisés. Comme pour R_{12} , la contribution de G à la mesure de changement diminue lorsque le nombre de GRH augmente. En effet, un plus grand nombre de GRH cible mieux la non-réponse, provoquant ainsi une plus grande variation dans le système de poids intermédiaires. Puisque pour l'ensemble des méthodes, G est constant, R_{int} est près de zéro et R_{12} est presque constant, il est clair que les variations de D sont influencées en majeure partie par les variations de R_{01} .

Méthode	D	R_{01} ($\times 10^{-3}$)	R_{01}/D (%)	R_{12} ($\times 10^{-3}$)	R_{12}/D (%)	R_{int} ($\times 10^{-5}$)	R_{int}/D (%)	G ($\times 10^{-2}$)	G/D (%)
1-GRH	0,012135	0,00	1,17	9,66	0,00	0,00	0,00	1,11	90,34
RL_4	0,012952	0,78	6,04	1,10	8,49	0,06	0,01	1,11	85,46
RL_16	0,013809	1,66	11,97	1,00	7,31	3,76	0,54	1,11	80,19
RL_40	0,014426	2,32	16,02	0,96	6,66	4,02	0,55	1,11	76,77
RL_60	0,014948	2,85	19,00	0,95	6,35	3,75	0,49	1,11	74,15
MS_16	0,015712	3,42	21,33	0,97	6,19	3,40	0,43	1,11	72,05
MS_25	0,016713	4,44	26,02	0,95	5,73	2,95	0,36	1,11	67,89
MS_40	0,018202	5,97	32,37	0,95	5,23	1,20	0,14	1,11	62,26

Tableau 1
Valeur moyenne de D sur les répétitions, de chaque composante et de leur contribution (en %) à la mesure de changement pour chacune des huit méthodes d'ajustement de la non-réponse

$$BRM = \frac{1}{M} \sum_{i=1}^M BR_i \text{ et } BAM = \frac{1}{M} \sum_{i=1}^M BA_i$$

où Y_i est l'estimation de la variable d'intérêt obtenue pour la i -ième répétition, $i = 1, 2, \dots, M$, $M=100$ pour la RL et $M=20$ pour le MS et Y est le total de la variable d'intérêt obtenu de l'échantillon de référence.

Le biais relatif moyen (BRM) et le biais absolu moyen (BAM) sont calculés en prenant respectivement la moyenne du BR et du BA pour l'ensemble des répétitions :

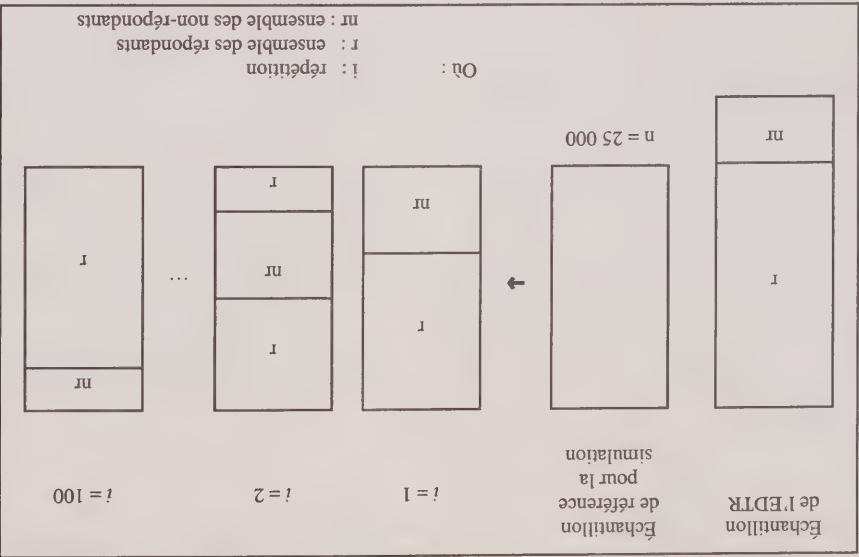
$$BR_i = 100 \left(\frac{Y_i - Y}{Y} \right) \text{ et } BA_i = Y_i - Y;$$

Afin de comparer la performance de la RL face au MS dans la réduction du biais de non-réponse, le biais relatif (BR) et le biais absolu (BA) ont été utilisés :

5.2.2 Biais relatifs et absolus

Le graphique 2 permet de comparer la RL et le MS en ce qui a trait à la contribution moyenne en pourcentage de R_{01} à D . Pour un nombre de GRH donné, R_{01} contribue à un plus grand pourcentage de D pour la méthode du MS que pour la RL. Ceci signifie que les changements individuels dans les poids entre le système initial et le système intermédiaire sont supérieurs pour le MS que pour la RL.

Le graphique 1 montre la contribution moyenne en pourcentage de R_{01} et R_{12} à la mesure de changement. Pour la RL et le MS, la contribution de R_{01} augmente avec le nombre de GRH tandis que celle de R_{12} diminue. Aussi, la contribution de R_{01} est supérieure pour le MS que pour la RL, tandis que celle de R_{12} est inférieure pour le MS que pour la RL. En fait, le profil de la contribution de R_{01} est le même que le profil de D (tableau 1). Ceci confirme que les variations de la mesure de changement sont principalement causées par les variations de R_{01} .



Plusieurs variantes des méthodes de sélection des variables ont été étudiées :

a) RL_{-i} , où i représente la moyenne approximative, sur les 100 répétitions, du nombre de GRH générés selon la méthode de la RL_{-40} , dans la présente étude, $i=4, 16, 40, 60$. Par exemple, pour RL_{-40} , les $q=6$ variables explicatives les plus importantes de la propension à répondre ont d'abord été identifiées. Les GRH ont ensuite été formés en utilisant les combinaisons valides ($2^q - j$) de ces $q=6$ variables explicatives. L'imposition de contraintes additionnelles ($n > 30$ et $TR > 50\%$) dans chaque GRH a mené au regroupement de certains GRH. En moyenne, sur les 100 répétitions, 24 GRH ont dû être regroupés ($j=24$) et un total de $2^q - j = 2^6 - 24 = 40$ GRH ont été formés, d'où l'appellation RL_{-40} . Dans l'étude de simulation, RL_{-i} , où $i=4, 16, 40, 60$ GRH, correspond respectivement à $q=2, 4, 6, 8$ variables explicatives.

b) MS_{-i} , où i indique la moyenne approximative sur les 20 premières répétitions du nombre de GRH générés selon la méthode du MS_{-16} , dans la présente étude, $i=16, 25, 40$. Par exemple, pour MS_{-16} , un MS a été utilisé avec un seuil de signification de 0,0001. Après l'imposition des mêmes contraintes additionnelles que pour la RL_{-16} , 16 GRH ont en moyenne été créés. MS_{-i} , où $i=16, 25, 40$ GRH, correspond respectivement aux seuils de signification 0,0001; 0,0005; 0,0025. Plus le seuil utilisé est élevé, plus il est facile d'identifier des différences significatives, ce qui permet une segmentation plus détaillée et en conséquence un plus grand nombre de GRH.

Le tableau 1 présente la valeur moyenne de D et des composantes sur toutes les M répétitions (où $M=100$ pour la RL et $M=20$ pour le MS) ainsi que la contribution en pourcentage de chaque composante à la valeur moyenne de D . On remarque tout d'abord que pour la méthode $1_GRH, R_{01}$ est nul puisqu'un seul et même ajustement de non-réponse est apporté à l'ensemble des répondants. Ainsi, $w_{1k} = \alpha w_{0k}$

5.2 Analyse des résultats de l'étude empirique

Pour chacune des méthodes discutées à la section précédente, les composantes de la mesure de changement D ont été étudiées. De même, le biais de non-réponse moyen, absolu et relatif ainsi que la variance moyenne des estimations ont été analysés.

5.2.1 Mesure de changement (D)

Pour chacune des méthodes discutées à la section précédente, les composantes de la mesure de changement D ont été étudiées. De même, le biais de non-réponse moyen, absolu et relatif ainsi que la variance moyenne des estimations ont été analysés.

réponse uniforme à l'intérieur de chacun des GRH. Ainsi, le facteur d'ajustement pour la non-réponse est donné par l'inverse du taux de réponse (pondéré par w_k ou non-pondéré) du GRH.

5. ÉTUDE EMPIRIQUE BASÉE SUR L'ENQUÊTE SUR LA DYNAMIQUE DU TRAVAIL ET DU REVENU (EDTR)

Afin de comparer l'efficacité de la RL et du MS, les données de l'EDTR ont été utilisées pour une étude empirique. L'EDTR est une enquête longitudinale auprès des ménages qui a débuté en 1993, et dont l'un des objectifs consiste à comprendre le bien-être économique de la société canadienne (voir Lavigne et Michaud 1998).

Ces deux méthodes ont été mises à l'essai à l'aide d'une simulation en analysant des variables d'intérêt et différents domaines. Les composantes de la mesure de changement, les biais absolus et relatifs de même que les variantes, ont été étudiés.

5.1 Description de l'étude empirique

Comme première étape de l'étude empirique, la probabilité de répondre à la première vague de l'enquête a été estimée pour chacune des unités de l'échantillon longitudinal. Des variables potentiellement explicatives de

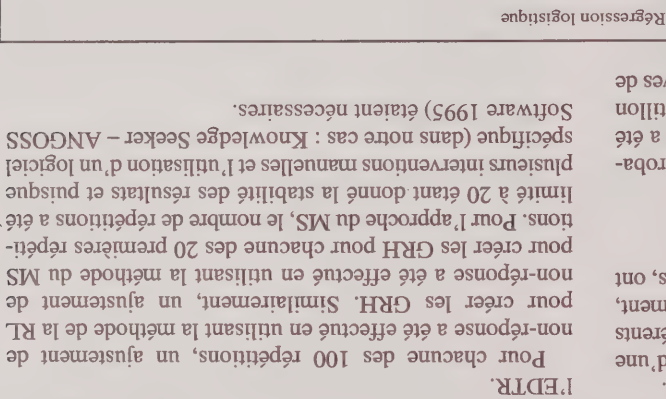


Figure 1. Visualisation de la formation des GRH selon la méthode

4. STRATÉGIES D'AJUSTEMENT DE LA NON-RÉPONSE

Dufour, Gagnon, Morin, Renaud et Sarnadal : Mieux comprendre la transformation des poids

Dans la littérature, il existe plusieurs méthodes d'ajuste-

ment de poids (dont la méthode décrite à la section 2.2) afin de compenser pour la non-réponse. Une autre méthode fréquemment utilisée dans les enquêtes longitudinales consiste à ajuster les poids selon l'inverse de la probabilité prédite de réponse obtenue à l'aide de la régression logistique. On retrouve également les méthodes d'ajustement basées sur le calage aux marges qui utilisent les distributions marginales de l'échantillon initial ou de la population. Singh, Wu et Boyer (1995) ont utilisé cette approche afin de dériver une méthode d'ajustement ayant la propriété de produire des estimations cohérentes d'une vague à l'autre pour les enquêtes longitudinales. Deville (1998) propose une méthode de correction de non-réponse par calage ou par échantillonnage équilibré. Pour une revue des méthodes d'ajustement de non-réponse, se référer à Kalton et Kasprzyk (1986), Platek, Singh et Tremblay (1978), Chapman, Bailey et Kasprzyk (1986) ainsi qu'à Little (1986). Dans le présent document, seules les méthodes reposant sur la création de GRH sont considérées.

4.1 Formation de GRH

Dans la plupart des enquêtes, l'information connue pour les non-répondants est minimale, mis à part quelques variables de stratification de la base de sondage. Ainsi donc, le choix des GRH est très limité et les strates sont souvent utilisées comme GRH. Dans un tel cas, l'hypothèse que la probabilité de réponse est la même pour toutes les unités d'une strate donnée est émise. Toutefois, dans les enquêtes longitudinales, beaucoup d'information à propos des répondants et des non-répondants de la vague courante est disponible à partir des réponses aux vagues précédentes. Cette quantité d'information peut donc être utilisée pour créer des GRH pour lesquels l'hypothèse d'un mécanisme de réponse uniforme à l'intérieur de ceux-ci est plausible. Ceci résulte en un meilleur ajustement pour la non-réponse et conséquemment en une réduction du risque d'introduction d'un biais de non-réponse dans les estimations.

4.1.1 Méthode de sélection des variables pour la formation des GRH

Par définition, les GRH sont formés à partir d'un ensemble de variables pouvant prédire la propension à répondre. Si l'ensemble de variables initialement identifiées est trop vaste, des tests univariés peuvent être utilisés afin de garder les variables les plus importantes pour discriminer les caractéristiques des répondants de celles des non-répondants. À partir de cet ensemble de variables importantes, une méthode de sélection peut être utilisée pour garder les meilleures variables explicatives de réponse. Deux méthodes courantes de sélection de

variables sont : un modèle de régression logistique (RL) et un modèle de segmentation (MS).

4.1.1.1 Régression logistique

Dans la méthode de la RL, l'utilisation combinée du « fait d'avoir répondu ou non à l'enquête » comme variable dépendante, des poids standardisés et de la procédure « stepwise », permet d'obtenir la liste des variables dichotomiques les plus significatives pour expliquer la propension à répondre. Règle générale, les GRH sont créés selon les 2^q combinaisons possibles à partir d'un ensemble de q variables explicatives retenues. On qualifie souvent la méthode de la RL d'approche symétrique. Cependant, le respect de certaines contraintes additionnelles lors de la création des GRH, peut entraîner une diminution de leur nombre. Par exemple, on peut exiger d'avoir un nombre minimum (n) d'unités de référence et un taux de réponse (TR) (pondéré ou non pondéré) supérieur à un certain seuil dans chacun des GRH. Kalton et Kasprzyk (1986) favorisent ces contraintes afin d'éviter l'augmentation de la variance associée aux poids extrêmes. Cependant, ces contraintes peuvent réduire l'efficacité de l'ajustement de non-réponse et se traduire par une augmentation du biais. Lorsqu'un GRH ne satisfait pas à une de ces contraintes, on doit le regrouper avec un autre GRH. Le regroupement des GRH s'effectue jusqu'à ce que tous les GRH satisfassent aux contraintes additionnelles imposées. De cette façon, on obtient $2^q - j$ combinaisons valides, où j représente la réduction causée par le regroupement des GRH.

Par exemple, à la figure 1, $2^4 = 8$ GRH sont créés à partir de $q=3$ variables explicatives. Les cases ombrées de la figure 1 représentent les GRH. À l'intérieur de chaque GRH, un facteur d'ajustement est calculé et le poids w_0^* de chaque unité de référence est ensuite ajusté correctement.

4.1.1.2 Modèle de segmentation

La méthode MS, dite non symétrique, est basée sur l'algorithme CHAD (Chi-square Automatic Interaction Detection) développé par Kass (1980). Elle partitionne l'échantillon en sous-groupes selon les taux de réponse des variables explicatives en se servant de tests du Khi-deux. La procédure de segmentation continue jusqu'à ce qu'une variable explicative significative ne puisse plus être trouvée. Les sous-groupes finaux créés par le MS deviennent les GRH, pour lesquels des ajustements de non-réponse sont calculés. Comme pour la RL, le respect de contraintes additionnelles peut être imposé.

À la figure 1, on peut observer que la méthode du MS a partitionné l'échantillon en plusieurs GRH à partir de différentes variables explicatives. Les GRH sont à nouveau représentés par les cases ombrées. La segmentation s'est poursuivie jusqu'à ce qu'il ne soit plus possible de trouver de variables explicatives.

4.1.2 Facteur d'ajustement pour la non-réponse

Que les GRH soient formés en faisant appel à la RL ou à l'aide du MS, on émet l'hypothèse d'un mécanisme de

Le ratio \bar{w}_{01} mesure le changement moyen du système de poids intermédiaires par rapport au système de poids initiaux. Plus la non-réponse totale est prononcée, plus \bar{w}_{01} s'éloigne de la valeur 1 qui est obtenue seulement en l'absence de non-réponse. Le ratio \bar{w}_{02} représente le changement moyen du système de poids finaux par rapport au système de poids initiaux.

Les ratios \bar{w}_{01} et \bar{w}_{02} mesurent un changement de poids moyen. Pour mesurer un changement de poids individuel, on définit, pour tout $k \in r$, $r_{01k} = w_{1k}/(w_{0k}\bar{w}_{01})$, et $r_{02k} = w_{2k}/(w_{0k}\bar{w}_{02})$. Ces quantités varient autour de 1. Plus précisément, leurs moyennes pondérées sont égales à 1 :

$$\frac{\sum_r w_{0k} r_{01k}}{\sum_r w_{0k} r_{02k}} = \frac{\sum_r w_{0k}}{\sum_r w_{0k}} = 1.$$

Les quantités r_{01k} et r_{02k} seront utiles pour mesurer les changements de poids individuels.

Le changement total que subissent les poids, du système de poids initiaux au système de poids finaux en passant par le système de poids intermédiaires, peut être calculé par une mesure de changement, aussi appelée *distance*. Soit D la mesure de changement suivante :

$$D = \frac{\sum_r w_{0k}}{\sum_r w_{0k} \left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2}.$$

En fait, D est une moyenne pondérée des facteurs individuels de changement de poids suivants :

$$\left(\frac{w_{2k}}{w_{0k}} - 1 \right)^2 = \left(\frac{w_{2k}}{w_{0k}} \frac{w_{1k}}{w_{0k}} - 1 \right)^2.$$

La mesure de changement D se décompose en quatre composantes selon l'équation suivante :

$$D = R_{01} + R_{12} + R_{int} + G$$

où :

$$R_{01} = \frac{\sum_r w_{0k} (r_{01k} - 1)^2}{\sum_r w_{0k}},$$

$$R_{12} = \frac{\sum_r w_{0k} (r_{02k} - r_{01k})^2}{\sum_r w_{0k}},$$

$$R_{int} = 2 \frac{\sum_r w_{0k} (r_{01k} - 1)(r_{02k} - r_{01k})}{\sum_r w_{0k}}$$

$$G = (\bar{w}_{02} - 1)^2.$$

Il est à noter que la mesure de changement D prend uniquement des valeurs positives, avec l'égalité à zéro quand les deux conditions suivantes sont satisfaites :

- (i) absence de non-réponse ($r = s$ et $w_{1k} = w_{0k}$ pour tout k),
- (ii) absence d'effet de la poststratification sur les poids intermédiaires ($w_{2k} = w_{1k}$ pour tout k).

Un taux de non-réponse élevé aura tendance à augmenter la valeur de la mesure de changement D étant donné que, dans un tel cas, w_{1k} est en moyenne considérablement plus grand que w_{0k} .

Le terme R_{01} mesure les changements individuels que subissent les poids en passant du système initial au système intermédiaire. On verra un peu plus loin que la composante R_{01} est en quelque sorte associée à la qualité du modèle pour la non-réponse et qu'une grande valeur de R_{01} est préférable. Le terme R_{12} mesure les changements individuels que subissent les poids en passant du système intermédiaire au système final. Le terme R_{int} mesure l'interaction entre les deux types de changements et le terme G mesure le changement de poids moyen entre le système initial et le système final.

En plus de son interprétation comme une distance, la mesure de changement D peut aussi être interprétée comme un écart quadratique moyen des changements w_{2k}/w_{0k} par rapport à 1, et ceci relativement à la distribution définie par les w_{0k} . Dans cette optique, la composante G correspond au biais au carré (soit le carré de l'écart entre la moyenne \bar{w}_{02} des w_{2k}/w_{0k} et 1) tandis que la somme des trois autres composantes correspond à la variance. Dans le plus simple des cas, où un ajustement de non-réponse est calculé en utilisant un seul GRH et où on n'applique aucune post-stratification, on a $w_{0k} = N/n$ pour tout $k \in s$ (dans le cas d'un tirage aléatoire simple de taille n) et $w_{1k} = w_{2k} = N/n$ pour tout $k \in r$ (où le facteur d'ajustement de non-réponse est n/m , soit l'inverse du taux de réponse). On a alors $D = G = \{(n/m) - 1\}^2$ et $R_{01} = R_{12} = R_{int} = 0$.

D'importantes conclusions peuvent être tirées en regardant l'importance relative des termes R_{01} , R_{12} et R_{int} . Si R_{01} est élevé, et qu'en même temps R_{12} est peu élevé, l'enquête en est une où l'ajustement de non-réponse cause d'importants changements individuels dans les poids, tandis que la poststratification ne modifie les poids individuels que très peu. Par contre, lorsque R_{12} est élevé, la poststratification engendre des changements individuels considérables. Les résultats présentés aux sections 5 et 6 démontreront que le terme R_{01} peut être utilisé dans la comparaison de l'efficacité de diverses méthodes d'ajustement de non-réponse. De plus, le signe de R_{int} indique si les deux types de changements individuels opèrent dans la même direction ($R_{int} > 0$) ou dans des directions opposées ($R_{int} < 0$). En pratique, on s'attend à ce que R_{int} soit peu important ou même négligeable.

non-réponse rencontrées dans la littérature. Viennent, aux sections 5 et 6, les résultats des études basées sur l'EDTR et l'ENLÉJ. La dernière section présente les conclusions de cette étude.

2. CADRE GÉNÉRAL POUR LA PONDERATION LONGITUDINALE

Dans une enquête longitudinale auprès des ménages, les personnes composant l'échantillon initial sont suivies à travers le temps et sont communément appelées *personnes longitudinales*. C'est cet ensemble de personnes qui sera utilisé dans les études présentées dans ce document. On référera à celles-ci en utilisant le terme « *unité de référence* ». La présente section donne un aperçu des étapes suivies afin de modifier le poids initial des personnes longitudinales en un poids final.

2.1 Poids initiaux

Soit $U = \{1, \dots, k, \dots, N\}$ une population finie. On s'intéresse à une variable y dont la valeur pour la k -ième unité est notée y_k . L'objectif est d'estimer le total $Y = \sum y_k$. Soit w_{0k} , le poids initial pour toute unité $k \in s$, où s désigne l'échantillon longitudinal. En l'absence de non-réponse, le système de poids initiaux $\{w_{0k} : k \in s\}$ donnerait un estimateur $\hat{Y} = \sum_s w_{0k} y_k$ pour Y . On suppose ici que les w_{0k} sont normés de façon à ce que $\sum_s w_{0k} = N$. Quoique sans biais pour Y , \hat{Y} a l'inconvénient de ne pas incorporer d'information auxiliaire sous la forme de totaux de contrôle connus pour des poststrates.

2.2 Ajustement de non-réponse et poids intermédiaires

La plupart des enquêtes font face à de la non-réponse. Deux approches sont souvent utilisées afin de compenser pour la non-réponse : l'imputation et la correction des poids initiaux des répondants par un facteur d'ajustement. Cette dernière est la plus couramment utilisée dans les enquêtes auprès des ménages pour pallier la non-réponse totale, tandis que l'imputation est souvent privilégiée dans le traitement de la non-réponse partielle. La non-réponse totale réduit la taille d'échantillon étant donné que la valeur y_k n'est disponible que pour $k \in r$, où $r \subset s$ est l'ensemble des m unités répondantes. Pour cet ensemble réduit de données, les poids initiaux w_{0k} sont en moyenne trop petits et on a $\sum_r w_{0k} < N$. L'estimateur $\hat{Y}' = \sum_r w_{0k} y_k$ n'est pas admissible car il sous-estime systématiquement Y .

L'ajustement des poids est souvent choisi afin de compenser pour la non-réponse totale dans les enquêtes auprès des ménages. Une méthode courante pour ajuster les poids consiste à construire des groupes de réponse homogènes (GRH). Ces derniers sont formés de sorte que chacun d'entre eux soit constitué d'unités de référence ayant une probabilité de réponse semblable. Ensuite, à l'intérieur de chaque GRH, un facteur d'ajustement, égal à l'inverse du

2.3 Poststratification et poids finaux

Une pratique courante dans les enquêtes auprès des ménages consiste à modifier les poids intermédiaires à l'aide d'une poststratification, ou de façon plus générale à l'aide d'un calage aux marges, de sorte que la somme des poids finaux sur l'ensemble des répondants corresponde aux comptes de population connus. La poststratification produit donc un système de poids finaux $\{w_{2k} : k \in r\}$, qui incorpore l'information auxiliaire et qui est à la fois cohérent avec les totaux de contrôle des poststrates. Dans ce cas, les poids finaux dans chaque poststrate p vérifient $\sum_p w_{2k} = N^p$ où N^p est l'effectif connu et r^p est l'ensemble des unités répondantes de la p -ième poststrate. Il s'ensuit que $\sum_r w_{2k} = N$. Des variables démographiques ou géographiques sont fréquemment utilisées pour définir des poststrates. Le choix des poststrates, qui doivent être suffisamment grandes, est limité par la disponibilité des totaux de contrôle. Plusieurs méthodes peuvent être utilisées afin de caler les poids intermédiaires aux totaux de contrôle choisis.

3. MESURE DE CHANGEMENT DES POIDS INITIAUX AUX POIDS FINAUX

Dans cette section, une mesure du changement entre les poids initiaux et les poids finaux est proposée afin de mieux comprendre l'effet de la procédure de modification de poids. La décomposition de cette mesure en quatre composantes permet de quantifier l'effet de chacune des étapes de pondération décrites à la section 2. Ces composantes seront utilisées aux sections 5 et 6 dans la comparaison de diverses méthodes d'ajustement des poids pour pallier la non-réponse.

Si les poids initiaux sont normés tels que $\sum_s w_{0k} = N$, et si $r \subset s$, alors les trois systèmes de poids décrits à la section 2 vérifient les relations suivantes :

Soit

$$\sum_r w_{0k} < N, \sum_r w_{1k} = N, \sum_r w_{2k} = N.$$

$$\bar{w}_{01} = \frac{\sum_r w_{1k}}{\sum_r w_{0k}} \text{ et } \bar{w}_{02} = \frac{\sum_r w_{2k}}{\sum_r w_{0k}}.$$

Mieux comprendre la transformation des poids à l'aide d'une mesure de changement

JOHANE DUFOUR, FRANÇOIS GAGNON, YVES MORIN, MARTIN RENAUD et CARL-ERIK SÄRNDAL¹

RÉSUMÉ

La littérature concernant les enquêtes longitudinales auprès des ménages propose plusieurs approches pour créer un ensemble de poids finaux à utiliser lors des analyses de données. La plupart de ces approches font appel à plusieurs procédures pour modifier les poids. Dans les faits, l'ensemble de poids initiaux est souvent transformé en un ensemble de poids intermédiaires afin de compenser pour la non-réponse, et par la suite en un ensemble de poids finaux via la poststratification pour redresser l'échantillon. La littérature dédie un grand intérêt à cette démarche mais aucune étude ne s'est vraiment penchée sur une approche pour mesurer l'importance relative de ces deux étapes ou pour mesurer l'efficacité des nombreuses alternatives qui existent pour créer les poids intermédiaires. L'objectif de cet article consiste à étudier et à mesurer la transformation des poids finaux (du poids initial au poids final) qui est produite par la procédure de modification des poids. Une décomposition des poids finaux est proposée pour évaluer l'ajustement de non-réponse, la correction pour la poststratification et l'interaction entre ces deux ajustements. On utilise cette mesure de changement comme outil pour comparer l'efficacité de diverses méthodes d'ajustement pour la non-réponse, notamment les méthodes reposant sur la formation de groupes de réponses homogènes. La mesure de changement est étudiée par l'entremise d'une étude de simulation utilisant les données d'une enquête longitudinale de Statistique Canada, soit l'Enquête sur la dynamique du travail et du revenu. La mesure de changement est également appliquée aux données d'une deuxième enquête longitudinale, c'est-à-dire à l'Enquête nationale longitudinale sur les enfants et les jeunes.

MOTS CLÉS : Non-réponse; pondération; calage; enquête longitudinale; mesure de changement.

1. INTRODUCTION

Il est courant de retrouver dans la littérature une démarche de transformation de poids en deux étapes pour des enquêtes auprès des ménages. Comme première étape, un ajustement est appliqué aux *poids initiaux* afin de compenser pour la non-réponse; les poids résultants sont appelés les *poids intermédiaires*. La deuxième étape produit les *poids finaux* en faisant appel à la poststratification ou plus généralement à la technique de calage (voir Deville et Särndal 1992) afin de s'assurer que les poids finaux respectent certains totaux de contrôle de population connus. Toutes ces modifications de poids sont effectuées avec comme objectif la production du « meilleur ensemble possible de poids finaux ».

À Statistique Canada, les enquêtes longitudinales auprès des ménages suivent également cette démarche en deux étapes lors de la pondération et les travaux de recherche entrepris par l'Agence penchent dans cette direction. Le « Survey of Income and Program Participation (SIPP) » du U.S. Bureau of the Census (voir Rizzo, Kalton et Brick 1996) suit également une telle approche.

Dans la littérature, plusieurs méthodes d'ajustement des poids sont proposées pour pallier la non-réponse. Rizzo et coll. (1996) ont comparé les estimations obtenues de plusieurs de ces méthodes à des estimations de sources indépendantes. Cependant, peu d'auteurs ont effectué des études de simulation ou proposé des outils pour comparer

L'efficacité relative des méthodes quant à leur capacité à réduire le biais de non-réponse.

L'objectif principal de ce document consiste à étudier et à mesurer le changement (entre les poids initiaux et les poids finaux) qui est produit suite à l'adoption d'une procédure de modification de poids en deux étapes. Pour ce faire, une mesure de changement à quatre composantes est proposée pour quantifier l'incidence relative de l'ajustement de non-réponse, la correction pour la poststratification et l'interaction entre ces deux ajustements. Le second objectif est d'utiliser la mesure de changement afin de comparer l'efficacité de diverses méthodes d'ajustement de non-réponse à l'aide d'une étude de simulation basée sur les données de l'Enquête longitudinale sur la dynamique du travail et du revenu (EDTR) et avec les données de l'Enquête nationale longitudinale sur les enfants et les jeunes (ENLEJ). Les enquêtes longitudinales se distinguent puisque beaucoup d'information à propos des répondants et des non-répondants aux vagues précédentes. Ceci permet l'utilisation de méthodes plus complexes pour l'ajustement de non-réponse.

Un cadre général pour la pondération des enquêtes longitudinales auprès des ménages est tout d'abord présenté à la section 2. Suit ensuite, à la section 3, une présentation de la mesure de changement qui sera utilisée pour quantifier les étapes de transformation entre le poids initial et le poids final. La section 4 traite de stratégies d'ajustement pour la

¹ Johane Dufour, François Gagnon, Yves Morin, Martin Renaud et Carl-Erik Särndal, Statistique Canada, Parc Tunney, Ottawa, (Ontario) K1A 0T6.

- RUBIN, D.B. (1987a). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley and Sons.
- RUBIN, D.B. (1987b). The SIR-algorithm - A discussion of Tanner and Wong's. The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91, 473-489.
- RUBIN, D.B., et SCHAFER, J.L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceeding of the Statistical Computing Section of the American Statistical Association*, 83-88.
- SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data by Simulation*. New York : Chapman and Hall.
- SILVERMAN, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of Royal Statistical Society, B*, 47, 1-52.
- SISCOVICK, D.S., RAGHUNATHAN, T.E., KING, I., WEINMANN, S., WICKLUND, K.G., ALBRIGT, J., BOVBERG, V., ARBOGAST, P., KUSHI, L., COBB, L., COPASS, M.K., PSATY, B.M., RETZLAFF, B., CHILDS, M. et KNOPP, R.H. (1995). Dietary intake and cell-membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. *Journal of American Medical Association*, 274, 1363-1367.

- Variable de comptage :** Pour X_j , une variable de comptage, ajuster un modèle de régression de Poisson $X \sim \text{Poisson}(\lambda)$, où $\log \lambda = U/\beta$. Les imputations pour des valeurs manquantes en X sont créées à l'aide des étapes ci-dessous :
1. Soit B_j , l'estimation correspondant à un maximum de vraisemblance de β_j , V sa matrice de covariances et T la décomposition de Cholesky de V . Produire un vecteur z d'écarts aléatoires normaux de lignes de dimension (B) et définir $\beta_j^* = B + Tz$.
 2. Soit U_j^{miss} , la portion de U pour laquelle X manque. Définir $\lambda_j^* = \exp(U_j^{\text{miss}} \beta_j^*)$. Produire des variables aléatoires de Poisson indépendantes avec des moyennes comme éléments de λ_j^* .
- Variable polynomique :** Pour X pouvant avoir valeurs $j = 1, 2, \dots, k$, noter $\pi_j = \Pr(X = j|U)$. Ajuster un modèle de régression polynomique établissant le lien entre X et U , où $\log = (\pi_j/\pi_k) = U/\beta_j$, pour $j = 1, 2, \dots, k-1$. Compte tenu de la restriction $\sum_{j=1}^k \pi_j = 1$, il s'ensuit que $\pi_k = (1 + \sum_{j=1}^{k-1} \exp(U/\beta_j))^{-1}$.

Soit B_j , l'estimation correspondant à un maximum de vraisemblance des coefficients de régression $(\beta_1^*, \beta_2^*, \dots, \beta_{k-1}^*)$, V la matrice des covariances asymptotique et T sa décomposition de Cholesky.

Les étapes ci-dessous permettent de créer des imputations :

 1. Définir $\beta_j^* = B + Tz$, où z est un vecteur d'écarts aléatoires normaux de lignes de dimension (B) .
 2. Soit U_j^{miss} , la ligne de U comportant des X manquants; soit $P_j^* = \exp\{\sum_{j=1}^k U_j^{\text{miss}} \beta_j^*\} / \{1 + \sum_{j=1}^k \exp(U_j^{\text{miss}} \beta_j^*)\}$, où β_j^* , représente les éléments appropriés de β_j^* , où $j = 1, 2, \dots, k-1$ et $P_k^* = 1 - \sum_{j=1}^{k-1} P_j^*$.
 3. Soit $R_0 = 0$, $R_j = \sum_{i=1}^j P_i^*$ et $R_k = 1$, les sommes cumulatives des probabilités. Pour imputer des valeurs, produire un nombre aléatoire uniforme u et considérer j comme la catégorie imputée si $R_{j-1} < u \leq R_j$.

Encore une fois, l'imputation de variables mixtes, de comptages et catégories provient de distributions prédictees à posteriori approximatves puisque les paramètres correspondants sont tirés de leurs distributions à posteriori approximatves normales asymptotiques.

BIBLIOGRAPHIE

BARNARD, J. 1995. Cross-procedures for Multiple Imputation Inference: Bayesian Theory and Frequentist Evaluation. Thèse Doctorat, non publiée, University of Chicago, Department of Statistics.

GELFAND, A.E., et SMITH, A.M.F. (1990). Sampling based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398-409.

GELMAN, A., CARLIN, J., STERN, H. et RUBIN, D.B. (1995). *Bayesian Data Analysis*. London, Chapman and Hall.

GELMAN, A., et SPEED T.P. (1993). Characterizing a joint probability distribution by conditionals. *Journal of Royal Statistical Society, B*, 55, 185-188.

GEMAN, S., et GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

HEERINGA, S.G., LITTLE, R.J.A. et RAGHUNATHAN, T.E. (1997). Imputation of Multivariate Data on Household Net Worth. University of Michigan, Ann Arbor, Michigan.

LI, K.H., MENG, X.L., RAGHUNATHAN, T.E. et RUBIN, D.B. (1991). Significance levels from repeated p values from multiply-imputed data. *Statistical Sinica*, 1, 65-92.

LI, K.H., RAGHUNATHAN, T.E. et RUBIN, D.B. (1991). Large sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of American Statistical Association*, 86, 1065-1073.

LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York : Wiley.

LITTLE, R.J.A., et SCHLUCHTER, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 72, 497-512.

LIU, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of multivariate analysis*, 53, 139-158.

MADOW, W.G., NISSELSOHN, H., OLKIN, I. et RUBIN, D.B. (1983). *Incomplete Data in Sample Surveys*. 1, 2, et 3, New York, Academic Press.

MENG, X.L., et RUBIN, D.B. (1992). Performing likelihood ratio tests with multiply imputed data sets. *Biometrika*, 79, 103-111.

OLKIN, I., et TATE, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*, 32, 448-465.

RAGHUNATHAN, T.E., et GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of American Statistical Association*, 90, 54-63.

RAGHUNATHAN, T.E., et RUBIN, D.B. (1988). An application of Bayesian statistics using sampling/importance resampling to a deceptively simple problem in quality control. *Data Quality Control : Theory and Pragmatics*, (G.E. Liepins et V.R.R. Uppluri, eds). New York: Marcel Dekker.

RAGHUNATHAN, T.E., et SISCOVICK, D.S. (1996). A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

RUBIN, D.B. (1976). Inference and missing data (avec discussion). *Biometrika*, 63, 581-592.

RUBIN, D.B. (1978). Multiple imputation in sample surveys - A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

de façon à obtenir l'estimation ponctuelle et sa matrice de covariances. Cette stratégie exige plus de calculs que la stratégie d'imputation multiple. Cette stratégie d'imputation JRR intégrée et plusieurs de ses variantes sont actuellement à l'étude.

Enfin, il a été supposé que le fichier de données provient d'un plan de sondage aléatoire simple. Toutefois, la plupart des enquêtes ont recours à des plans de sondage complexes mettant en jeu la stratification, le groupement et la pondération. Il y a lieu de poursuivre les travaux afin de modifier la méthode de régression séquentielle de façon à incorporer des fonctions de plan complexe qui ne sont pas reflétées dans les variables X de l'expression (1). Toutefois, même si le processus d'imputation ne tient pas compte des fonctions de plan complexe, l'analyse des données complètes devrait se fonder sur le plan. Même si cela ne donne pas des inférences fondées sur le plan qui sont valides, la robustesse que sous-tend l'analyse fondée sur le plan est conservée dans un certain mesure. La stratégie d'imputation JRR intégrée dont il est question ci-dessus pourrait offrir des propriétés fondées sur le plan plus attrayantes dans le cadre d'un plan complexe.

REMERCIEMENTS

Les auteurs aimeraient remercier les trois arbitres pour l'attention particulière qu'ils ont apportée à la lecture de cet article ainsi que pour leurs remarques pertinentes. Cette recherche a été partiellement soutenue par une subvention NSF DMS-0803720.

ANNEXE : MODÈLES DE RÉGRESSION ET IMPUTATIONS

Si l'on abandonne, par souci de brièveté, les indices inférieurs des variables, les étapes nécessaires à l'imputation de chaque type de variable sont les suivantes :

Variable continue : Pour X (possiblement transformée à partir de l'échelle originale pour la normalité), une variable continue, construire un modèle de régression linéaire normale, $Y = U\beta + e$, où U est la matrice des variables explicatives mises à jour le plus récemment, e comporte une distribution normale multivariée avec une moyenne zéro et une variance $\sigma^2 I$, et I est une matrice d'identité. Supposer que $\theta = (\beta, \log \sigma)$ comporte une distribution a priori uniforme sur l'espace réel dimensionnel approprié. Ajuster ce modèle en fonction des unités pour lesquelles X est observée.

Soit $B = (U'U)^{-1}U'Y$, le coefficient de régression estimé, $SSE = (Y - UB)'(Y - UB)$, la somme des carrés des résidus et df = lignes (X) - cols (U) , les degrés de liberté des résidus, et T la décomposition de Cholesky telle que $TT' = (U'U)^{-1}$. Il est facile de calculer les distributions a posteriori pertinentes (voir, par exemple, Gelman,

Carlin, Stern et Rubin 1995, chapitre 7); les étapes ci-dessous fournissent ensuite des tirages de la distribution prédictive a posteriori des valeurs X manquantes :

1. Produire un écart aléatoire chi carré u avec df degrés de liberté, et définir $\sigma_u^2 = SSE/u$.
2. Produire un vecteur $z = (z_1, z_2, \dots, z_p)$ de dimension p = lignes (B) d'écart aléatoires normaux, et définir $\beta_* = B + \sigma_u Tz$.
3. Soit U_{miss} , la matrice U pour celles qui ont des valeurs X manquantes. Les valeurs imputées sont $X_* = U_{miss}\beta_* + \sigma_u v$, où v est un vecteur indépendant de lignes de dimension (U_{miss}) d'écart aléatoires normaux.

Variable binaire : Lorsque X est une variable binaire, on ajuste un modèle de régression logistique établissant le lien entre X et U (mises à jour le plus récemment), $\logit[\Pr(X = 1 | U)] = U\beta$, à l'aide d'unités pour lesquelles X est observée. Les valeurs imputées pour X sont créées suivant les étapes ci-dessous :

2. Soit B , les estimations correspondant à un maximum de vraisemblance de β , et V sa matrice de covariances asymptotiques (inverse négatif de la matrice d'informations de Fisher observées). Soit T , la décomposition de Cholesky de V (c'est-à-dire, $TT' = V$). Produire un vecteur z d'écart aléatoires normaux des lignes de dimension (B) . Définir $\beta_* = B + Tz$.

2. Soit U_{miss} , la portion de U pour laquelle X manque. Définir $P_* = [I + \exp(-U_{miss}\beta_*)]^{-1}$. Produire un vecteur u , de lignes de dimension (U_{miss}) de nombres aléatoires uniformes entre 0 et 1. Imputer 1 si une composante particulière de u est inférieure ou égale à la composante correspondante de P_* , et imputer 0 autrement.

Cette stratégie ne donne que des tirages approximatifs de la distribution prédictive a posteriori des valeurs manquantes car les tirages du paramètre β proviennent de l'approximation asymptotique de sa distribution a posteriori réelle. Il est possible de puiser dans la distribution réelle en modifiant l'étape 1 à l'aide, par exemple, de l'échantillonnage-importance-rééchantillonnage (Rubin 1987b).

Variable mixte : Pour X , une variable mixte (c'est-à-dire que X a une valeur soit nulle, soit continue), modéliser les valeurs nulles à l'aide d'un indicateur 0-1 afin de distinguer entre 0 et des valeurs autres que 0, puis modéliser une variable à distribution normale pour le volet continu de la distribution à la condition que la variable indicatrice soit égale à 1. Autrement dit, utiliser une stratégie à deux degrés: imputer 1 ou 0 à l'aide de la stratégie logistique décrite ci-dessus, puis, en limitant l'échantillon aux unités ayant une valeur non nulle, utiliser la stratégie de la variable continue décrite ci-dessus pour imputer une valeur continue remplaçant la valeur 1 qui vient d'être imputée.

oeuvre de cette procédure exige uniquement un bon gène-
rateur de nombres aléatoires et des programmes d'ajustage
pour un choix de programmes de régression multiple. Une
application à base de SAS permettant d'exécuter cette
stratégie peut être téléchargée d'un site Web
(www.isr.umich.edu/src/smp/iv/).
Dans certains cas, il est possible de modifier l'algorithme
de façon à en faire un échantillonnage de Gibbs à partir de
la distribution prédictive composée des valeurs manquantes
compte tenu des valeurs observées. Toutefois, la procédure
IMRS sera plus utile lorsqu'il est difficile de formuler un
modèle explicite. Tant pour les illustrations que pour la
simulation, différents points de départ aléatoires ont servi à
surveiller les valeurs imputées, aspect important dans de
nombreuses applications concrètes. Il s'agit d'une bonne
pratique lorsqu'un échantillonnage de Gibbs est utilisé dans
le cadre d'un modèle bayésien explicite (Gelman et Rubin
1992), et elle devait être utilisée lorsqu'on a recours à la
méthode de régression séquentielle décrite dans le présent
exposé.
L'étude de simulation décrite à la section 5, bien que
limitée, est favorable pour ce qui est des inférences fondées
sur l'IMRS. Les imputations relevant des modèles IMRS et
de Bayes étaient comparables. Il s'agissait ici, toutefois,
d'élaborer une stratégie d'imputation qui soit peu influée une
variable à la fois et complètement en fonction de toutes les
informations observées, plutôt qu'une distribution multi-
dimensionnelles composée explicite de toutes les variables.
De plus, on peut réduire la sensibilité du modèle en ayant
recours à un modèle de régression semi-paramétrique pour
chaque régression conditionnelle. L'interprétation bayé-
sienne des modèles de lissage de type spline (Silverman
1985) peut servir à tirer des valeurs imputées de la
distribution prédictive. De telles modifications méritent
également une recherche plus poussée.
Pour certains grands fichiers de données comportant de
nombreuses variables, l'IMRS peut exiger beaucoup de
temps d'ordinateur. On peut modifier l'algorithme de façon
à appliquer une méthode de sélection des variables pour
chaque régression de chaque cycle. Nous avons comparé les
inférences avec et sans la sélection de variables pour
plusieurs grands fichiers de données, par exemple la
National Health Interview Survey et la National Medical
Expenditure Survey, à l'aide de plusieurs centaines de
variables. Les inférences descriptives aussi bien que les
inférences fondées sur des modèles de régression linéaire et
logistique étaient très semblables, mais il subsiste un besoin
d'implémenter la stratégie d'imputation multiple des valeurs
manquantes.

Nous avons décrit et évalué une procédure d'imputation
multidimensionnelle par régression séquentielle pouvant
servir à imputer les valeurs manquantes d'un choix de
structures de données complexes comportant de nombreux
types de variables, de restrictions et de limites. Cette
procédure devrait être utile lorsqu'il est difficile de définir
une distribution composée de toutes les variables ayant des
valeurs manquantes. Un réel avantage de la stratégie est sa
souplesse lorsqu'il s'agit de traiter chaque variable indivi-
duellement. Ainsi, afin de conserver toutes les corrélations
entre deux variables, il faut inclure tous les termes à effet
majeur à titre de variables explicatives, et pour conserver,
par exemple, trois interactions factorielles, il faut inclure
toutes les interactions à deux facteurs à titre de variables
explicatives dans le modèle d'imputation. La mise en

6. DISCUSSION

Coefficients de régression	Moyenne	ET	IMRS	Données complètes
β_0	8,2	2,0	96,1	95,4
β_1	8,8	1,7	95,4	94,9
β_2	8,0	2,2	95,3	94,7

Tableau 4
Moyennes et écarts types de la différence normalisée entre les estimations IMRS et les estimations de données complètes d'une part et la couverture réelle d'intervalle de confiance de 95 % nominaux d'autre part

estimations d'intervalle bien calées.
1,22. Autrement dit, les données de l'IMRS ont donné des
de l'intervalle de confiance des données complètes pour β_1
était de 0,91 et pour l'IMRS la grandeur moyenne était de
1,22. Autrement dit, les données de l'IMRS ont donné des
estimations d'intervalle bien calées.
On a utilisé la même étude de simulation afin de com-
parer les propriétés distributives des imputations de l'IMRS
et d'une méthode entièrement bayésienne. Pour les hypo-
thèses modélisées servant à préparer des données com-
plètes, nous avons élaboré un algorithme de Monte Carlo à
chaînes markoviennes afin de tirer des valeurs de la distri-
bution prédictive à posteriori réelle des valeurs manquantes
compte tenu des valeurs observées. Chaque étape du tirage
faisait appel à l'algorithme de Metropolis-Hastings et
exigeait apparemment plus de temps d'ordinateur que la
méthode IMRS. Par conséquent, seuls les 500 premiers
fichiers de données simulées ont été utilisés pour cette
comparaison. Nous avons calculé deux statistiques de
Kolmogorov-Smirnov (K_S) à partir de chaque fichier de
données simulées : une pour comparer les imputations de la
méthode IMRS et les valeurs cachées réelles, l'autre pour
comparer les imputations bayésiennes et les valeurs cachées
réelles. Il n'y avait aucune différence discernable entre ces
deux statistiques pour les 500 fichiers de données simulées.
Un nuage de points autour d'une pente de 45 degrés.

5. ÉTUDE PAR SIMULATION

L'analyse décrite aux sections 3 et 4 indique que les analyses des résultats raisonnables en appliquant la stratégie IMRS au traitement des données manquantes. Néanmoins, il est difficile de déterminer, d'après de telles études de cas, si la stratégie donnera des inférences valides dans des applications courantes. Une étude de simulation a été conçue afin d'examiner les propriétés d'échantillonnage répétées des inférences tirées de fichiers de données imputées créés à l'aide de la stratégie IMRS. On a tiré de populations hypothétiques des fichiers de données complètes et on a supprimé des éléments en vertu d'un mécanisme de données manquantes ignorables. On a imputé les valeurs supprimées et on a évalué les différences de statistiques sommaires d'après les fichiers de données imputées et les fichiers de données antérieures à la suppression ou complètes.

Plus précisément, la stratégie a servi :

(1) à préparer un fichier de données complètes qui ne correspondait pas tout à fait à notre stratégie d'imputation multiple,

(2) à estimer des paramètres de régression choisis,

(3) à supprimer certaines valeurs à l'aide d'un mécanisme de données manquantes ignorables,

(4) à utiliser l'IMRS en vue de la polyimputation des valeurs manquantes, et

(5) à obtenir des estimations polyimputées pour les paramètres de régression estimés à l'étape 2.

Les différences du paramètre sont examinées pour plusieurs répétitions indépendantes de cette stratégie.

Au total, on a préparé 2 500 fichiers de données complètes comportant trois variables (U , X_1 , X_2) et une taille d'échantillon de 100 à l'aide des modèles ci-dessous :

1. $U \sim \text{Normal}(0, 1)$;

2. $X_1 \sim \text{Gamma}$ avec une moyenne $\mu_1 = \exp(U-1)$ et une variance $\mu_1^2/5$; et

3. $X_2 \sim \text{Gamma}$ avec une moyenne $\mu_2 = \exp(-1 + 0,5U + 0,5X_1)$ et une variance $\mu_2^2/2$.

Le modèle pour X_2 à l'étape 3 est le modèle de régression primaire d'intérêt avec de vrais coefficients de régression $\beta_0 = -1$, $\beta_1 = \beta_2 = 0,5$, et un paramètre de dispersion $\phi^2 = 0,5$. Pour les données complètes, on peut ajuster ce modèle à l'aide de logiciels statistiques comme GLIM ou SpIus.

(1) aucune valeur manquante en U ;

(2) des valeurs manquantes en X_1 qui dépendent de U suivant une fonction logistique logit [$\Pr(X_1 \text{ manquant}) = 1,5 + U$; et

(3) des valeurs manquantes en X_2 qui dépendent de U et de X_1 suivant une fonction logistique logit [$\Pr(X_2 \text{ manquant}) = 1,5 - 0,5X_1 - 0,5U$].

Ces mécanismes de données manquantes ont généré 22 % de données manquantes en X_1 et 29 % de données manquantes en X_2 . L'analyse de cas complets n'aurait utilisé que 48 % des données.

Puisque l'IMRS nous permet seulement d'ajuster un modèle de régression linéaire normale, les imputations ont été exécutées comme suit. Supposons que X_1 a moins de valeurs manquantes, et notons $Z_1 = (X_1^* - 1) / \lambda_1$, la transformation de Box-Cox de la variable continue. Dans le premier cycle d'imputations, supposons que Z_1 a une distribution normale avec une moyenne $a_0 + a_1U$ et une variance σ_1^2 , où l'on estime λ_1 à l'aide de la stratégie du maximum de vraisemblance, et que $Z_2 = ((\lambda_2^2 - 1) / \lambda_2)$ a une distribution normale avec une moyenne $b_0 + b_1U + b_2Z_1$ et une variance σ_2^2 , où l'on estime λ_2 à l'aide du maximum de vraisemblance. Pour les cycles subséquents, U et Z_2 sont des variables explicatives pour Z_1 , et U et Z_2 sont des variables explicatives pour Z_2 . L'estimation d'une transformation exponentielle à l'aide du maximum de vraisemblance a été automatisée au moment d'ajuster chaque modèle de régression.

Pour chacun des 2 500 fichiers de données simulées comportant des valeurs manquantes, on a créé au total 250 cycles ayant $M=5$ différents points de départ aléatoires à l'aide de l'IMRS. Pour chaque répétition, on a analysé les $M=5$ fichiers de données imputées résultants et le fichier de données complètes (avant la suppression) en ajustant le modèle Gamma pour X_2 à l'aide du maximum de vraisemblance. L'estimation polyimputée a été construite comme la moyenne des cinq estimations des données imputées. Afin d'évaluer les différences des estimations ponctuelles, nous avons calculé les différences normalisées entre l'IMRS et des estimations de données complètes,

$$\Delta(\beta) = 100 \times$$

$$\frac{\text{abs(estimation IMRS - estimation de données complètes)}}{\text{ET(estimation IMRS)}}.$$

Le tableau 4 indique la moyenne et l'écart type de $\Delta(\beta)$ pour trois coefficients de régression β_0 , β_1 et β_2 dans le modèle. Les estimations IMRS se situent typiquement à 8 % près des unités standard complètes. Pour les coefficients de régression, on a calculé la couverture réelle et la grandeur (1987b). Pour chaque fichier de données simulées et chaque paramètre, on a déterminé si la vraie valeur ($\beta_1 = 0,5$ par exemple) se trouve à l'intérieur de l'intervalle correspondant. On a calculé la proportion d'intervalle contenant les vraies valeurs pour les 2 500 répétitions (voir le tableau 4). Pour ce qui est des fichiers de données complètes, la couverture réelle pour β_1 , par exemple, était de 94,9 % et pour l'IMRS de 95,4. De plus, on a calculé la largeur moyenne des intervalles de confiance. La largeur moyenne

nous avons énoncé le modèle de régression à effets mixtes ci-dessous,

$$X_{2ic} = \alpha_0 + \alpha_1 U_{1i} + \alpha_2 U_{2i} + \alpha_3 X_{1ic} + \delta_i + \epsilon_{ic}.$$

où δ_i et ϵ_{ic} sont des variables aléatoires normales indépendantes l'une de l'autre avec une moyenne 0 et des variances σ_{δ}^2 et σ_{ϵ}^2 respectivement. Encore une fois, l'absence de données manquantes dans les covariables, il est facile d'obtenir les estimations correspondant à un maximum de vraisemblance des paramètres inconnus en utilisant, par exemple, la procédure PROC MIXED de SAS. Il n'y avait pas de valeurs manquantes dans la classification des groupes de risque, et nous avons donc défini $X = (1, U_1, U_2)$. Les variables comportant des valeurs manquantes, X_{21}, X_{22}, X_{31} et X_{32} ont été imputées par régression linéaire normale, et les valeurs manquantes en X_{11} et X_{12} ont été imputées par régression logistique. Nous avons créé $M=25$ IMRS, en répétant le processus pour 1 000 cycles et 25 points de départ différents. Les fichiers de données polyimputées IMRS ont été analysés et combinés à l'aide des méthodes décrites antérieurement. Afin de comparer ces résultats aux inférences polyimputées lorsque les imputations sont tirées de la distribution prédictive a posteriori relevant du modèle d'emplacement général, nous avons créé 25 imputations relevant d'un modèle entièrement bayésien à l'aide d'un logiciel préparé par Schafer (1997). Le tableau 3 contient les estimations ponctuelles et les erreurs types des trois modèles fondés sur stratégies d'imputation multiple IMRS et de Bayes. Il n'existe pas de différences réelles significatives entre les estimations et les erreurs types IMRS d'une part et celles qui résultent de l'imputation bayésienne d'autre part. Les enfants de parents du groupe à risque élevé ont environ 7,8 [exp (2,048)] fois plus de chances d'avoir un nombre élevé de symptômes que les enfants de parents du groupe normal dans le cadre de l'IMRS. L'intervalle de confiance de 95 % pour ce risque relatif est de (3,8, 16,0). Pour le groupe à

Tableau 3
Estimations ponctuelles (erreurs types) des coefficients de régression pour trois modèles de développement chez l'enfant dans le cadre d'une imputation IMRS et de Bayes

Variables explicatives		Méthode d'imputation		Symptômes		Note en lecture		Note en compréhension verbale	
Ordonnée à l'origine	IMRS	-0,678	(0,256)	4,654	(0,013)	4,873	(0,020)	4,991	(0,021)
	Bayes	-0,688	(0,257)	4,556	(0,013)	4,991	(0,021)	4,991	(0,021)
Groupe à risque élevé	IMRS	2,048	(0,356)	-0,109	(0,022)	-0,191	(0,032)	-0,191	(0,032)
	Bayes	2,033	(0,350)	-0,108	(0,021)	-0,180	(0,033)	-0,180	(0,033)
Groupe à risque modéré	IMRS	1,289	(0,366)	-0,110	(0,022)	-0,162	(0,033)	-0,162	(0,033)
	Bayes	1,300	(0,360)	-0,109	(0,023)	-0,167	(0,035)	-0,167	(0,035)
Symptômes	IMRS	-	-	0,032	(0,022)	-0,083	(0,032)	-0,083	(0,032)
	Bayes	-	-	0,031	(0,019)	-0,080	(0,030)	-0,080	(0,030)

variable dépendante et un certain nombre de variables observées complètement à titre de variables explicatives ont indiqué que les données ne manquent pas complètement au hasard. On peut donc s'attendre à ce que les estimations de cas complets et les erreurs types soient biaisées.

Le tableau 2 (IMRS, méthode 1) contient des estimations et leurs erreurs types pour l'IMRS d'après les variables du modèle de fond seulement. Ces estimations sont assez semblables à celles de l'analyse de cas complets. Les erreurs types de l'imputation multiple sont plus petites à cause des sujets additionnels ayant des données imputées. Il y a de faibles changements du rapport entre l'usage du tabac et l'arrêt cardiaque primaire. L'analyse de cas complets indique un rapport statistiquement significatif entre les années d'usage du tabac et l'arrêt cardiaque primaire pour des personnes qui ont déjà fumé, tandis qu'une telle association n'est pas révélée par l'analyse des données polyimputées.

Un des avantages de la stratégie d'imputation multiple est que le processus d'imputation peut faire appel à des variables additionnelles qui ne se trouvent pas dans l'analyse de fond. De telles situations se présentent lorsqu'une base de données de recherche commune comportant plusieurs variables est utilisée par différents chercheurs, ayant chacun recours à un sous-ensemble des variables. L'imputation peut se faire pour la base de données entière, la prédiction des valeurs manquantes pour chaque variable étant renforcée par toutes les autres variables du fichier de données. On a pu montrer que de telles imputations améliorent l'efficacité comparative à celles qui se fondent uniquement sur les variables d'un modèle de fond partiel (Raghuathan et Siscock 1996).

Le tableau 2 (IMRS, méthode 2) contient des estimations de l'imputation multiple et leurs erreurs types obtenues lorsque le fichier de données entier a été imputé en fonction de 50 variables additionnelles. Celles-ci comprenaient des indicateurs socioéconomiques et des variables de comportements. Les estimations ponctuelles sont légèrement différentes pour toutes les variables. Les erreurs types, par contre, sont appréciablement plus petites comparativement à la stratégie d'imputation fondée uniquement sur des variables du modèle de fond (IMRS, méthode 1). Il n'y a là rien de surprenant car plusieurs des variables additionnelles, par exemple la tension artérielle, le compte de cholestérol, la consommation d'alcool et l'activité physique, étaient fortement prédictives de l'indice de masse corporelle et des variables liées à l'usage du tabac.

4. TROUBLES PSYCHOLOGIQUES PARENTAUX ET DÉVELOPPEMENT CHEZ L'ENFANT

Une deuxième illustration examine les effets des troubles psychologiques parentaux sur plusieurs mesures du développement chez l'enfant. Little et Schuchter (1985) ont

analysé les données à l'aide d'un modèle d'emplacement général afin d'obtenir des estimations correspondant à un maximum de vraisemblance des paramètres de la distribution composée. On a utilisé ce modèle d'emplacement des méthodes de Monte Carlo à chaînes markoviennes (Scharf 1997), ce qui a donné des fichiers de données polyimputées entièrement à base de modèle bayésien. Nous avons également créé des imputations multiples à l'aide de la procédure IMRS.

Les données de l'étude se rapportent à 69 familles ayant chacune deux enfants. Chaque famille a été classée dans une des trois catégories de risque suivantes : 1) risque normal ; aucun trouble psychiatrique parental ; 2) risque modéré ; diagnostic chez un parent d'un trouble psychiatrique ou d'une maladie physique chronique ; 3) risque élevé : diagnostic chez un parent de schizophrénie ou de trouble mental affectif. Il y avait trois variables dépendantes primaires d'intérêt : X_{1c} , nombre de symptômes psychiatriques (dichotomisation : élevé/faible) chez l'enfant c ; X_{2c} , test de lecture normalisé chez l'enfant c ; et X_{3c} , test normalisé de compréhension verbale chez l'enfant c .

Nous considérons trois modèles pour l'étude de l'effet des troubles psychologiques parentaux sur le développement des enfants. Le premier est un modèle de régression logistique à effets mixtes :

$$\logit[\Pr(X_{1c} = 1)] = \beta_0 + \beta_1 U_{1c} + \beta_2 U_{2c} + \gamma_{1c}$$

où $X_{1c} = 1$ si un enfant c d'une famille i est considéré comme ayant un nombre élevé de symptômes et 0 autrement; $U_{1c} = 1$ si une famille i est considérée comme étant exposée à un risque modéré et 0 autrement; $U_{2c} = 1$ si une famille i est considérée comme étant exposée à un risque élevé et 0 autrement; et γ_{1c} sont des effets aléatoires considérés comme étant des variables aléatoires normales distribuées de façon identique et indépendante avec une moyenne 0 et une variance ϕ_1^2 . Cet effet aléatoire tient compte d'une corrélation intraclasses entre deux enfants au sein d'une même famille. Pour des données complètes, on peut ajuster ce modèle en maximisant la fonction de vraisemblance numériquement intégrée de $(\beta_0, \beta_1, \beta_2, \phi_1^2)$ à l'aide de l'algorithme de Newton-Raphson et de la méthode de quadrature gaussienne pour l'intégration numérique de la fonction de vraisemblance. Il est facile d'ajuster ce type de modèle avec des données complètes, mais difficile de le faire avec des valeurs manquantes.

Les deuxième et troisième modèles de régression établis sent le rapport entre les notes de l'enfant en lecture et en compréhension verbale, respectivement, pour un enfant c d'une famille i , logarithme des notes en lecture et en compréhension verbale, respectivement. Ainsi, en notant X_{2c} et X_{3c} comme miqu était appropriée. Une étude des valeurs résiduelles après quelques cycles préliminaires ou imputations de notes en lecture et en compréhension verbale a indiqué qu'une échelle logarithmique était appropriée. Ainsi, en notant X_{2c} et X_{3c} comme

$\hat{\alpha}^{(t)}$ est l'estimation du vecteur de coefficients de régression en fonction du fichier de données imputées l . L'estimation α du modèle logistique, et $V^{(t)}$ sa matrice des covariances, polyimputée de α est

$$\hat{\alpha}_{MI} = \sum_{M=1}^l \hat{\alpha}^{(t)} / M$$

et sa matrice des covariances est

$$V_{MI} = \sum_{M=1}^l V^{(t)} / M + \frac{M+1}{M} B_M$$

où

$$B_M = \sum_{M=1}^l (\hat{\alpha}^{(t)} - \hat{\alpha}_{MI})(\hat{\alpha}^{(t)} - \hat{\alpha}_{MI})' / (M - 1)$$

Le nombre d'imputations dépasse la valeur recommandée. Nous avons exécuté 25 imputations avec différents points de départ aléatoires afin de déterminer si les cycles de type Gibbs mènent à une région des valeurs imputées qui est très différente des valeurs observées. Des affichages graphiques des valeurs imputées et observées ont indiqué qu'aucune des imputations des 25 000 cycles n'était incompatible avec la distribution des données observées. Le tableau 2, l'analyse de cas complets, contient les estimations ponctuelles et leurs erreurs types fondées sur 103 sujets (11,5 %) avaient des valeurs manquantes pour une ou plusieurs variables explicatives. Une analyse de cas complets, qui n'est généralement valide que lorsque les données sont manquantes tout à fait au hasard, a été exécutée après l'élimination de ces 103 sujets (voir la colonne 2, tableau 2). Des analyses de régression logistique comportant un indicateur de données manquantes à titre de

Tableau 2
Estimations ponctuelles (erreurs types) des coefficients de régression logistique pour le modèle de l'arrêt cardiaque primaire pour des cas complets, méthodes IMRS 1 et 2**

Variables explicatives		Cas complets		IMRS	
		(n = 795)		Méthode 1 (n = 898)	Méthode 2 (n = 898)
		Estimation (ET)		Estimation (ET)	Estimation (ET)
Ordonnée à l'origine		-2,922	(0,791)	-2,610	(0,757)
Âge		0,015	(0,009)	0,015	(0,009)
Femme		-0,007	(0,203)	-0,115	(0,189)
Éducation		-0,448	(0,173)	-0,467	(0,166)
Indice de masse corporelle		0,056	(0,018)	0,049	(0,013)
Personne qui fume		1,693	(0,569)	2,001	(0,543)
Personne qui a fumé		0,003	(0,284)	-0,029	(0,262)
Personne qui fume × Années d'usage		-0,003	(0,015)	-0,008	(0,013)
Personne qui a fumé × Années d'usage		0,019	(0,009)	0,014	(0,009)
Personne qui a fumé × Années d'usage					0,014
					(0,009)

* Méthode 1 – Imputation limitée à des variables de modèle
** Méthode 2 – Imputation comprenant des variables modélisées et auxiliaires

Il existe d'autres cas particuliers dans lesquels cette approximation est l'équivalent du tirage de valeurs d'une distribution prédictive a posteriori relevant d'un modèle complètement paramétrique. Si donc toutes les variables sont continues et si chaque modèle de régression conditionnelle est un modèle de régression linéaire normale à variance constante, il y a convergence de l'algorithme vers une distribution prédictive composée relevant d'une distribution normale multidimensionnelle comportant une distribution a priori irrégulière pour la moyenne et la matrice des covariances.

Il est théoriquement possible qu'une série de tirages fondés sur les densités en (3) ne converge pas vers une distribution stationnaire, car ces densités conditionnelles ne sont peut-être compatibles avec aucune distribution conditionnelle composée multidimensionnelle de X_1, X_2, \dots, X_k étant donné X (Gelman et Speed 1993). Nos études empiriques fondées sur plusieurs fichiers de données concrètes n'ont toujours pas permis d'identifier ce genre d'anomalie. Dans plusieurs grands fichiers de données, les densités conditionnelles (2) et (3) paraissent assez semblables. Comme il a été mentionné aux sections 4 et 5, les tirages axés sur cette stratégie sont comparables à ceux qui se fondent sur un modèle bayésien explicite.

3. EFFET DE L'USAGE DU TABAC SUR L'ARRÊT CARDIAQUE PRIMAIRE

Dans notre première illustration, la stratégie IMRS est appliquée à une étude de cas-témoins portant sur la relation entre l'usage de la cigarette et l'incidence de l'arrêt cardiaque primaire (Siscovick, Raghunathan, King, Weimann, Wicklund, Albright, Bovbjerg, Arbogast, Kushi, Cobb, Copass, Psatsy, Retzlaff, Childs and Knopp 1995). Dans cette étude, il est difficile de formuler un modèle explicite qui englobe toute la complexité des données. Les sujets de cas étaient tous des résidents de King County, Washington, ayant subi un arrêt cardiaque primaire à l'extérieur de l'hôpital entre 1988 et 1994. Les sujets de cas ont été identifiés à l'aide d'un examen des rapports d'incidents paramédicaux. Les sujets témoins ont été sélectionnés à l'aide d'un sondage téléphonique au hasard (King County) et appartés à des sujets de cas en fonction du sexe et de l'âge (à sept ans près). Pour être admissibles, les sujets (cas et témoins) devaient être âgés de 25 à 74 ans, mariés et libres de toute maladie du coeur (diagnostic clinique) ou de tout autre péril comme un cancer, une maladie du foie, une maladie des poumons, ou encore une insuffisance rénale terminale.

Puisque l'arrêt cardiaque primaire comporte un taux de létalité supérieur à 80 %, le fait d'être marié a été ajouté comme critère d'admissibilité afin que l'information sur l'exposition au facteur de risque (état de fumeur, années d'usage) puisse être confirmée par les répondants substitués (conjoint). Parmi les sujets témoins et les sujets de cas de

où C est un indicateur de l'arrêt cardiaque. Les résultats préliminaires indiquent que des termes linéaires pour l'âge et l'indice de masse corporelle sont appropriés.

Tableau 1
Moyennes et proportions (en %) des variables clés et pourcentage manquant

Variable	Témoins ($n = 551$)	Cas ($n = 347$)
% manquant	% manquant	% manquant
Âge	58,4 (10,4)	59,4 (9,9)
Indice de masse corporelle	25,8 (4,1)	26,4 (4,6)
Années d'usage du tabac	24,8 (14,7)	31,7 (13,8)
Femme	23,2	19,9
≥ École sec.	76,8	61,9
État de fumeur	47,2	27,3
N'a jamais fumé	0,0	0,0
A déjà fumé	42,1	38,2
Fumeur actuellement	10,7	34,5

Il n'y a pas de valeurs manquantes pour les variables Âge, Femme, Éducation, État de fumeur (X_1, X_2) et C . Ainsi, pour ce qui est de l'imputation, définissons $X = (1, \text{Âge, Femme, Éducation, } X_1, X_2, C)$. Log (BMI), avec le moins de valeurs manquantes, a d'abord subi une régression sur X en fonction d'un modèle de régression linéaire normale. Des diagnostics résiduels ont indiqué qu'une transformation logarithmique améliorerait la normalité des résidus.

l'échantillon. À titre de covariable, cette variable pourra être traitée différemment lors de l'imputation de variables subséquentes. Ainsi, certaines variables fictives pourront être créées en fonction de cette variable, puis annexées à la matrice U avant que l'imputation de la

variable suivante ne se poursuive.

Considérons un autre exemple, « années d'usage de la cigarette », l'échantillon se limitant à des personnes qui fument ou qui ont fumé. En l'absence d'indication que ces personnes ont fumé au cours de leur adolescence, « années d'usage de la cigarette » devra pour une personne qui fume actuellement satisfaire la limite (0, Âge - 18). S'il y a lieu de croire que la personne a fumé au cours de son adolescence, on pourra restreindre l'étendue, par exemple : (0, Âge - 12). Pour une personne qui a déjà fumé, ces étendues seront (0, Âge - 18 - YRSQUIT) et (0, Âge - 12 - YRSQUIT), respectivement, où YRSQUIT représente le nombre d'années qui se sont écoulées depuis que la personne a cessé de fumer. Le modèle de régression approprié pour cette variable est une version tronquée du modèle de régression linéaire normale (possiblement en fonction d'une échelle transformée). Les paramètres, les coefficients de régression et la variance résiduelle doivent être tirés des distributions a posteriori correspondantes. Les imputations sont alors tirées de la distribution normale tronquée correspondante en fonction de la valeur tirée des paramètres.

Il est difficile de tirer des valeurs de paramètres directement de leur distribution a posteriori selon des vraisemblances normales tronquées. Toutefois, le calcul est facile pour une valeur de paramètre donnée. L'algorithme SIR (échantillonnage-importance-rééchantillonnage) (Rubin 1987b; Raghnathan et Rubin 1988) permet de puiser dans la distribution a posteriori elle-même. Tout d'abord, on tire plusieurs valeurs de paramètre d'essai de la distribution a posteriori sans appliquer les limites (modèle de régression linéaire normale non tronquée). Deuxièmement, on rattache un coefficient d'importance à chaque valeur d'essai, défini comme le rapport entre la densité a posteriori réelle avec limites et la densité d'essai (la densité a posteriori sans limites), les deux étant évaluées à la valeur tirée. Enfin, on échantillonne de nouveau une valeur de paramètre avec probabilité proportionnelle aux coefficients d'importance. Cette méthode exige une surveillance soignée de la distribution des coefficients d'importance (Gelman, Carlin, Stern et Rubin 1995).

Les limites s'appliquent également à des variables polytomiques. Ainsi, supposons qu'une variable X puisse avoir une valeur k quelconque, mais que les données observées indiquent que la valeur manquante pour un sujet particulier peut être soit l . Le rôle de ce sujet dans la vraisemblance correspond à la distribution binomiale conditionnelle. Les tirages de l'étape multinomiale (voir l'annexe A) se font à partir de la distribution conditionnelle pour ces deux catégories. Autrement dit, la valeur imputée est f avec probabilités $s_j = P_{fj} / (P_{fj} + P_{lj})$, et l avec la probabilité $1 - s_j$.

À la fin du premier cycle d'imputations, on a le premier fichier de données complètes sans valeurs manquantes. La factorisation à l'équation (1) définit une distribution conditionnelle composée de X_1, X_2, \dots, X_k ; étant donné X . Si le profil des données manquantes est monotone, les imputations du premier cycle sont des tirages approximatifs de la densité prédictive a posteriori composée des valeurs manquantes compte tenu des valeurs observées. À noter que les tirages des variables logistiques, polynomiques et de composition proviennent d'approximations (pour de grands échantillons) de la densité a posteriori des coefficients de régressions. Il est possible d'améliorer ces approximations en ayant recours, par exemple, à l'algorithme SIR ou à un autre algorithme de rejet à chaque cycle subséquent.

Lorsque le profil des données manquantes n'est pas monotone, il est possible d'élaborer un algorithme d'échantillonnage de Gibbs (Geman et Geman 1984; Gelfand et Smith 1990) qui correspond au modèle (1). Ainsi, moyennant les valeurs tirées des paramètres $\theta_1, \theta_2, \dots, \theta_k$ et les valeurs manquantes tirées du premier cycle, le deuxième cycle tirera des valeurs de θ_1 de la densité a posteriori conditionnelle appropriée qui est proportionnelle au premier terme de l'équation (1). Il s'agit ensuite de tirer les valeurs manquantes en X_1 moyennant cette valeur tirée du paramètre θ_1 , toutes les autres valeurs observées ou imputées pour ce sujet et d'autres paramètres $\theta_2, \theta_3, \dots, \theta_k$ du modèle. Autrement dit, les valeurs manquantes en X_j au cycle $(t + 1)$ doivent être tirées de la densité conditionnelle

$$f_j^*(X_j | \theta_1^{(t+1)}, X_1^{(t+1)}, \dots, \theta_k^{(t)}, X_k^{(t)}), \quad (2)$$

calculée en fonction de la distribution composée en (1), où $X_j^{(t)}$ représente les valeurs imputées ou observées pour la variable X_j au cycle t . Bien que cela soit conceptuellement possible, il est difficile de calculer même cette densité dans la plupart des situations concrètes vu les restrictions, les limites et le type de variables à l'étude.

Nous proposons un tirage des valeurs manquantes en X_j au cycle $(t + 1)$ à partir d'une distribution prédictive correspondant à la densité conditionnelle,

$$g_j(X_j | X_1^{(t+1)}, X_2^{(t+1)}, \dots, X_k^{(t)}, X_k^{(t)}), \quad (3)$$

où la densité conditionnelle g_j est définie par l'un des modèles de régression décrits antérieurement qui dépend du type de variable pour X_j , et ϕ_j représente les paramètres de régression inconnus ayant une distribution a priori diffusée. Autrement dit, les nouvelles valeurs imputées pour une variable dépendent des valeurs imputées antérieurement pour d'autres variables, et des valeurs nouvellement imputées de cette proposition peut être considérée comme une approximation d'un échantillonnage réel de Gibbs où la densité conditionnelle (3) fournit une approximation de la densité conditionnelle (2). De plus, on peut améliorer cette approximation en considérant l'algorithme SIR ou un autre algorithme de type rejet si la densité conditionnelle en (2) peut être calculée jusqu'à une constante.

modifier pour certaines distributions a priori convenables. Chaque régression conditionnelle se fonde sur un des modèles ci-dessous :

1. un modèle de régression linéaire normale à une échelle appropriée (par exemple, une transformée exponentielle de Box-Cox peut servir à atteindre la normalité) si X_j est continue;
2. un modèle de régression logistique si X_j est binaire;
3. un modèle de régression logit polytomique ou généralisée si X_j est catégorique;
4. un modèle linéaire logarithmique de Poisson si X_j est une variable de comptage;
5. un modèle à deux degrés dont la valeur nulle-non nulle est imputée par régression logistique; pour une valeur non nulle, un modèle de régression linéaire normale sert à imputer des valeurs non nulles, si X_j est composée.

Chaque imputation est constituée de c « cycles ». On commence le cycle 1 par régression de la variable comportant le plus petit nombre de valeurs manquantes, X_1 sur X , par imputation des valeurs manquantes en fonction du modèle de régression approprié. Si l'on suppose une distribution a priori plate pour les coefficients de régression, les imputations, pour les valeurs manquantes en X_1 sont les tirages de la distribution prédictive a posteriori correspondante (on trouve à l'annexe A des détails sur le tirage de valeurs pour divers modèles de régression). Il s'agit alors de mettre X à jour en annexant X_1 de façon appropriée (par exemple des variables fictives, si elle est catégorique) et de passer à la prochaine variable, X_2 , occupant le rang suivant parmi les valeurs manquantes les moins nombreuses. Le processus d'imputation est répété à l'aide de X mises à jour à titre de variables explicatives jusqu'à ce que toutes les variables aient été imputées. Autrement dit, il y a régression de X_1 sur U ; de X_2 sur $U=(X_1, X_1)$ où X_1 comporte des valeurs imputées; de X_3 sur $U=(X_1, X_1, X_2)$ où X_1 et X_2 comportent des valeurs imputées, et ainsi de suite.

La procédure décrite ci-dessus doit être modifiée si l'on veut incorporer des restrictions et des limites. Les sous-ensembles appropriés d'unités. Par exemple, un modèle de régression de Poisson peut être appliqué à l'imputation de valeurs manquantes pour la variable « nombre de grossesses ». L'imputation se limitera aux femmes de

cause des liens systématiques complexes entre les variables et les restrictions. Pour la deuxième application, on peut utiliser un modèle d'emplacement général pour créer des imputations multiples (Ojkin et Tate 1961 et Little et Schluchter 1985). Nous comparons donc des inférences d'imputations multiples résultant de la stratégie DMRS à des inférences résultant d'un modèle multidimensionnel composé. La section 5 contient les résultats d'une étude de simulation portant sur les propriétés d'échantillonnage d'inférences tirées de données imputées. Pour terminer, nous discutons de l'orientation des recherches à venir à la section 6.

2. MÉTHODE D'IMPUTATION

Pour un échantillon de taille n , nous notons X un plan ou une matrice explicative $n \times p$ contenant toutes les variables n'ayant pas de valeurs manquantes. X comporte des variables continues, binaires, de type comptage ou mixtes, ainsi que des variables fictives appropriées représentant des variables catégoriques. De plus, X peut comporter une colonne de uns afin de modéliser un paramètre de coordonnées à l'origine, des variables de décalage et certaines variables de plan. Soit X_1, X_2, \dots, X_k des variables k comportant des valeurs manquantes, ordonnées, en toute généralité, selon le nombre de valeurs manquantes, des moins nombreuses aux plus nombreuses. Le schéma n'est pas nécessairement monotonique. (Dans un schéma monotonique de valeurs manquantes, X_2 est observée uniquement pour un sous-ensemble de ceux pour lesquels X_2 est observée, et ainsi de suite.)

Pour des imputations modelisées, la densité conditionnelle composée de X_1, X_2, \dots, X_k compte tenu de X peut être factorisée comme suit

$$f(X_1, X_2, \dots, X_k | X, \theta_1, \theta_2, \dots, \theta_k) = f_1(X_1 | X, \theta_1) f_2(X_2 | X, X_1, \theta_2) \dots f_k(X_k | X, X_1, X_2, \dots, X_{k-1}, \theta_k) \quad (1)$$

où $f_j, j=1, 2, \dots, k$ sont les fonctions de densité conditionnelle et θ_j est un vecteur de paramètres de la distribution conditionnelle (par exemple, coefficients de régression et paramètres de dispersion). Dans le contexte d'une enquête sur échantillon, on peut considérer cela comme un modèle de superpopulation. Nous modelisons chaque densité conditionnelle à l'aide d'un modèle de régression approprié comportant des paramètres inconnus, θ_j , et nous puisons dans la distribution prédictive correspondante des valeurs manquantes comme tenu des valeurs observées. Nous supposons que la distribution a priori pour les paramètres $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ est $\pi(\theta) \propto 1$ (diffuse relativement à la vraisemblance). Toutefois, la méthode se laisse facilement

caractéristiques comme le fait d'avoir fumé au cours de l'adolescence. Dans le cas d'une personne qui a déjà fumé, x englobe également les années écoulées depuis la cessation d'usage du tabac. Un autre exemple de limites est abordé dans Heerhag, Little et Ragunathan (1997). Ces auteurs examinent l'imputation de questions comportant des réponses entre crochets lorsque le répondant ne peut pas ou ne veut pas donner une réponse exacte (au sujet du revenu ou de l'avis, par exemple), tout en définissant les limites à l'intérieur desquelles les valeurs imputées doivent se situer. Le présent exposé a comme objectif de proposer et d'évaluer une procédure d'imputation multidimensionnelle générale permettant de traiter une structure de données relativement complexe lorsque des modèles multidimensionnels complets explicites ne se laissent pas facilement former, les valeurs imputées pour chaque unité étant toutefois entièrement liées à toutes les valeurs observées pour l'unité en question. La stratégie consiste à considérer l'imputation, une variable à la fois, mais en fonction de toutes les variables observées. La stratégie de base permet de créer des imputations en vertu d'une série de régressions multiples, le type de modèle de régression variant selon le type de variable imputée. Les covariables englobant toutes les autres variables observées ou imputées pour l'unité en question. Les imputations sont définies comme des tirages de la distribution prédictive a posteriori exigée par le modèle de régression avec une distribution a priori plate ou non informative pour les paramètres du modèle de régression. La séquence d'imputation de valeurs manquantes peut se poursuivre d'une façon cyclique, se superposant à chaque fois aux valeurs tirées antérieurement, entraînant une interdépendance des valeurs imputées et misant sur la structure corrélatrice des covariables. Pour la création d'imputations multiples, on peut appliquer la même procédure avec différents points de départ aléatoires ou en prenant chaque P^e série de valeurs imputées des cycles mentionnés ci-dessus.

Les variables du fichier de données sont considérées comme relevant de l'un ou l'autre de cinq types : 1) continu, 2) binaire, 3) catégorique (polytomique avec plus de deux catégories), 4) de type comptage, 5) mixte (variable continue à masse de probabilité non nulle de 0). Du point de vue des calculs, les variables binaires et catégoriques se laissent traiter de façon identique, mais le fait de les distinguer facilite la conceptualisation et la description de l'algorithme de base. De plus, la population est considérée comme essentiellement infinie, l'échantillon étant simple et aléatoire et le mécanisme de données manquantes étant ignorable (Rubin 1976). Le recours à une imputation multiple en présence d'un plan complexe n'a toujours pas été étudié à fond et dépasse le cadre du présent exposé.

Nous décrivons ci-dessous la stratégie d'imputation multidimensionnelle par régression séquentielle (IMRS) à la section 2 et, aux sections 3 et 4, nous en évaluons deux applications. Il est difficile, pour la première application, de poser une distribution multidimensionnelle composée à

variables comportant des valeurs manquantes, lie aux variables observées intégralement et à certains paramètres inconnus, une distribution a priori pour les paramètres inconnus et un modèle du mécanisme des données manquantes, qu'il n'est pas nécessaire de préciser dans le cadre d'un schéma de données manquantes ignorables (Rubin 1976). Ce modèle explicite donne alors lieu à une distribution prédictive a posteriori pour l'imputation. Les imputations sont tirées de cette distribution prédictive a posteriori. Il existe plusieurs programmes et algorithmes informatiques pour l'imputation des valeurs manquantes en présence d'une normalité multidimensionnelle (Rubin et Schaffer 1990), de la distribution / multidimensionnelle (Liu 1995) et de diverses variations du modèle d'emplacement général (Schaffer 1997, Raghunathan et Grizzle 1995, Raghunathan et Siscovick 1996). Ce dernier modèle permet de traiter la distribution composée de variables catégoriques et continues; il a été abordé et proposé par Olkin et Tate (1961) et a été utilisé par Little et Schlichter (1985) en fonction précisément de problèmes de données manquantes. Une propriété importante de ces stratégies, c'est qu'elles dépendent entièrement de toute l'information observée. Plusieurs études de simulation (Raghunathan et Grizzle 1995, par exemple) indiquent que les inférences tirées de ce genre de données imputées offrent des propriétés d'échantillonnage souhaitables.

Les fichiers de données d'enquête comportent souvent de très nombreuses variables ayant différentes distributions. Typiquement, ces fichiers de données ont des centaines de variables, les uns continus, les autres de type comptage, souvent dichotomiques ou polytomiques, et même parfois dépendantes et semi-continues ou limitées. De plus, les variables continues peuvent comporter une distribution normale, normale logarithmique ou autre. Il peut être très difficile dans une telle situation de postuler un modèle bayésien intégral. De plus, les données d'enquête ont souvent deux autres caractéristiques qui rendent la modélisation encore plus complexe. Tout d'abord, certaines restrictions sont impératives. Ainsi, la variable « nombre d'années depuis la cessation de l'usage du tabac » est définie uniquement pour des personnes qui ont déjà fumé; par conséquent, le processus d'imputation pour cette variable devrait se limiter aux personnes qui ont déjà fumé. Certaines restrictions relatives d'instructions « passez à » dans un questionnaire. Ainsi, certaines questions sur le revenu d'un deuxième emploi sont posées uniquement lorsque le répondant indique qu'il ou elle a un deuxième emploi. L'imputation de ce genre de variable exige un traitement hiérarchique.

Deuxièmement, il existe des limites logiques ou des limites de cohérence pour les valeurs manquantes qu'il faut intégrer au processus d'imputation. Une telle interdépendance des variables rend la création du modèle difficile. Ainsi, « années d'usage du tabac » se limite aux personnes qui fument ou qui ont fumé, et les valeurs imputées doivent être inférieures à l'âge – x années, où x peut varier d'autres

Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression

TRIVELLORE E. RAGHUNATHAN, JAMES M. LEPKOWSKI, JOHN VAN HOEWYK
et PETER SOLENNBERGER¹

RÉSUMÉ

Le présent article décrit et évalue une procédure d'imputation des valeurs manquantes pour une structure relativement complexe des données lorsque celles-ci manquent au hasard. On obtient les imputations en ajustant une séquence de modèles de régression et en tirant les valeurs des distributions prédictives correspondantes. Les types de modèle de régression utilisés sont les suivants : linéaire, logarithmique, de Poisson, logit généralisé, ou encore un mélange qui dépend du type de variable imputé. Deux autres caractéristiques communes du processus d'imputation sont intégrées : la restriction à une sous-population pertinente pour certaines variables et des limites ou contraintes logiques pour les valeurs imputées. Les restrictions comportent la création de sous-ensembles d'unités d'échantillon répondant à certains critères au moment de l'ajustement des modèles de régression. Les limites supposent que l'on tire des valeurs d'une distribution prédictive tronquée. L'élaboration de cette méthode s'est inspirée en partie de l'analyse de deux fichiers de données utilisés à titre d'illustration. On applique la procédure de régression séquentielle à l'analyse d'imputations multiples pour les deux problèmes appliqués. Les propriétés d'échantillonnage des inférences tirées de fichiers de données polymérisées créés à l'aide de la méthode de régression séquentielle sont évaluées en fonction de fichiers de données simulées.

MOTS CLÉS : Non-réponse partielle; manquant au hasard; imputation multiple; mécanisme de données manquantes non-ignorable; régression; simulations et propriétés d'échantillonnage.

1. INTRODUCTION

Les données incomplètes sont un problème fréquent dans la plupart des recherches appliquées. On a élaboré plusieurs méthodes permettant de tirer des inférences de fichiers de données comportant des valeurs manquantes (Little et Rubin 1987), et ce travail se poursuit. Le schéma d'imputations multiples proposé par Rubin (1978, 1987a, 1996) est une possibilité intéressante si un fichier de données est destiné à plusieurs chercheurs ayant différentes compétences en statistique. Cette façon de procéder suppose l'imputation de plusieurs ensembles plausibles de valeurs manquantes dans le fichier de données incomplètes de façon à fournir plusieurs fichiers de données complètes. Chaque fichier de données complètes est analysé séparément, par exemple en ajustant un modèle de régression particulier. Les inférences qui en résultent (estimations ponctuelles et matrices de covariances) sont alors combinées à l'aide de la formule de Rubin (1987a, chapitre 3) et de perfectionnements de cette formule (Li, Raghunathan et Rubin 1991; Li, Meng, Raghunathan et Rubin 1992; et Barnard 1995).

De façon générale, les stratégies de traitement des données manquantes par imputation sont fort utiles dans la pratique, car une fois les valeurs manquantes imputées, on peut avoir recours à des logiciels de données complètes existants pour analyser les données. Puisque l'élaboration de logiciels pour l'analyse des données complètes évolue en bayésienne. Celle-ci exige un modèle explicite pour des données manquantes par imputation de méthodes d'imputation de divers points de vue à un long passé (Madaw, Nisselson, Olkin et Rubin 1983). Un schéma théoriquement élégant pour l'élaboration de méthodes d'imputation est la stratégie d'estimation de la variance. L'élaboration de méthodes d'imputation de la variance, autre procédure d'estimation de la variance, est également servie à créer une imputation unique avec une stratégie d'imputation décrite dans le présent exposé peut (Rao et Shao 1992), offrent également cet avantage. La technique de répétition répétée de type jackknife modifiée d'estimation de la variance appropriée, par exemple la méthode d'imputation unique en fonction d'une procédure comme l'imputation unique en fonction de données simulées et réelles, le bien-fondé de cette stratégie. D'autres possibilités montent, en analysant des fichiers de données simulées et pliquée cette technique dans différentes situations et ont (voir par exemple la bibliographie de Rubin 1996) ont ap- tons assez générales (Rubin 1987a). Plusieurs chercheurs estiment d'intervalles valides pour une série de conditions permet d'obtenir des estimations ponctuelles et des plètes permet d'obtenir des estimations ponctuelles et des avanage de la stratégie d'imputations multiples, c'est que nels peaufinés en présence d'un problème précis. Un autre quantes, seront en mesure d'ajuster des modèles fonction- en oeuvre de nouvelles procédures pour les données man- culières leur permettant de créer leur propre code de mise quées, sans connaître les ressources ou techniques appli- ques, les personnes qui s'adonnent à des recherches statisti- fonction de l'introduction de nouvelles méthodes statisti-

SCHAFER, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

SHAO, J., CHEN, Y. et CHEN, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

SHAO, J., et STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. *Journal of the American Statistical Association*, 94, 254-265.

B. Validité de l'équation (15) dans les conditions du mécanisme d'imputation sans remise

Nous supposons que $m = kr + t$ où k et t sont des nombres entiers non négatifs et que $t < r$. Supposons que l'estimateur de la moyenne de y a la forme (A.1). Supposons que l'imputation est exécutée de façon telle que t répondants sont utilisés $k + 1$ fois pour l'imputation et que $r - t$ unités sont utilisés k fois pour l'imputation. Les t répondants qui sont utilisés $k + 1$ fois sont choisis par échantillonnage aléatoire simple sans remise. Alors,

$$E_t(d_i) = k + r - 1 \quad t = r - 1, m$$

et

$$\text{Cov}_t(d_i, d_j) = \begin{cases} -r^{-1}t(1 - r^{-1}t) & \text{si } i = j \\ -r^{-2}t & \text{si } i \neq j. \end{cases}$$

Donc, par des arguments similaires à ceux utilisés dans la preuve de (A.2), nous obtenons

$$V(q_y) = V(\bar{y}_t) + E \left(n^{-2} r^{-1} t \sum_{i=1}^t \hat{e}_i^2 \right). \quad (\text{B.1})$$

Par conséquent, en utilisant (A.3) et (A.4), nous obtenons

$$V\{q_y\} = [n^{-1}R_2 + (r^{-1} + n^{-2}t)(1 - R_2)]\sigma_y^2. \quad (\text{B.2})$$

Maintenant, subordonnée à l'échantillon réalisé et aux répondants, nous obtenons

$$E_t \left\{ (1 + d_i)^2 \right\} = \left(\frac{r}{n} \right)^2 + \frac{r}{t} \left(1 - \frac{r}{t} \right)$$

de sorte que $V\{\mu_y\}$ dans l'équation (15) satisfait

$$E_t \left\{ V\{\mu_y\} \right\} = n^{-1}(n - 1) \sum_{i=1}^t (\hat{y}_i - \bar{y}_t)^2$$

$$+ [r^{-1} + n^{-2}t(1 - r^{-1}t)]$$

$$(r - p)^{-1} \sum_{i=1}^t (y_i - \hat{y}_i)^2.$$

Par conséquent, en utilisant (A.4) et (A.5), nous obtenons l'absence de biais approximative de $V\{\mu_y\}$ dans le cas du mécanisme d'imputation sans remise.

C. Preuve de l'équation (26)

Premièrement, définissons $R_n^{(t)} = (\bar{x}_1^{(t)} - \bar{x}_2^{(t)})(\hat{\beta} - \beta)$ et $R_n = (\bar{x}_1 - \bar{x}_2)(\hat{\beta} - \beta)$. D'après l'égalité (25),

$$V^* = \sum_{t=1}^T c_t^i \left(\bar{y}_t^{*(t)} - \bar{y}_t \right)^2 = A_n + B_n + 2C_n$$

BIBLIOGRAPHIE

- COCHRAN, W.G. (1977). *Sampling Techniques*. New York : John Wiley and Sons.
- FAY, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual Research conference*, 429-440.
- FAY, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 227-232.
- FAY, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.
- FULLER, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.
- HANSEN, M., HURWITZ, W.N. et MADOWS, W.G. (1953). *Sample Survey Methods and Theory*. Vol. I, New York : John Wiley and Sons.
- HANSEN, M., et TEPPING, B.J. (1985). Estimation for Variance in NAEP. Note de service inédite, Westat, Washington, D.C.
- RAO, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.
- RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- RAO, J.N.K., et SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley and Sons.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.
- SÄRNDAAL, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.
- SÄRNDAAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de la Statistique*, 55, 279-294.

Généralement parlant, les pseudo-données peuvent être décrites par la formule

$$y_i^* = \begin{cases} y_i' & i = r + 1, r + 2, \dots, n \\ y_i' + c_i g_i' (y_i' - y_i) & i = 1, 2, \dots, r \end{cases} \quad (39)$$

où y_i' est la valeur prévue de y_i dans les conditions du

modèle utilisé pour l'imputation. Si $c_i g_i' = 1$, alors l'estimateur de la variance traite les valeurs imputées comme des observations. En choisissant convenablement $c_i g_i' > 1$, on obtient un estimateur de la variance cohérent. Si la méthode d'imputation est déterministe et que les répondants utilisés sont considérés comme étant un échantillon aléatoire de l'échantillon original, alors $c_i' = r^{-1} n > 1$. Pour un échantillonage à deux phases avec plan de sondage complexe, $c_i = w_i' w_i^*$, où w_i' représente le poids d'échantillonage de l'unité i pour l'échantillon de première phase et w_i^* représente le poids d'échantillonage de l'unité i pour l'échantillon de seconde phase.

Le g_i de l'équation (39) représente la correction faite pour améliorer les propriétés conditionnelles, étant donné la variable auxiliaire x_i . Pour l'imputation par la méthode du quotient,

$$g_i' = (\bar{x}_2)^{-1} \bar{x}_1$$

où $\bar{x}_2 = \sum_{i=1}^n w_i' x_i^*$ et $\bar{x}_1 = \sum_{i=1}^n w_i' x_i$. Pour l'imputation par régression avec la grandeur scalaire x_i ,

$$g_i' = 1 + (\bar{x}_1 - \bar{x}_2) \left\{ \sum_{k=1}^K w_k^* (x_k - \bar{x}_2)^2 \right\}^{-1} (x_i - \bar{x}_2).$$

Dans l'un et l'autre cas, nous avons

$$\sum_{i=1}^n w_i' g_i' x_i = \bar{x}_1.$$

Pendant l'évaluation du présent article, Shao et Steel (1999) ont proposé des méthodes similaires dans le cas de l'imputation déterministe. Notre méthode est plus générale en ce sens que nous considérons aussi l'imputation aléatoire et que nous introduisons le terme c_i pour améliorer les propriétés de l'échantillon fini.

REMERCIEMENTS

L'auteur remercie son conseiller de thèse, Wayne A. Fuller, pour les discussions fructueuses qu'ils ont eues. Il remercie aussi Pamela Abbitt, F. Jay Breidt, Lou Rizzio, Richard Valliant et les examinateurs pour leurs commentaires utiles, qui lui ont permis d'améliorer considérablement l'article. La plupart des travaux présentés ici ont été réalisés pendant que l'auteur poursuivait des études de cycles supérieures à la Iowa State University et financés

ANNEXE

A. Preuve des équations (10) et (12)

L'estimateur $\hat{\mu}_{y_i}$ qui figure dans l'équation (9) peut être écrit sous la forme

$$\hat{\mu}_{y_i} = n^{-1} \sum_{i=1}^r \hat{y}_i' + n^{-1} \sum_{i=r+1}^n (1 + d_i') \hat{e}_i' \quad (A.1)$$

où d_i' est le nombre de fois que l'unité i est utilisée comme un donneur. En vertu du mécanisme d'imputation avec échantillonnage à probabilité égale et avec remise, nous avons

$$E_I(d_i') = r^{-1} m$$

et

$$\text{Cov}_I(d_i', d_j') = \begin{cases} -r^{-2} m & \text{si } i \neq j \\ r^{-1} m (1 - r^{-1}) & \text{si } i = j \end{cases}$$

où l'indice I représente la variation due au mécanisme d'imputation. Il s'ensuit que $E_I(\hat{\mu}_{y_i}) = n^{-1} \sum_{i=1}^n \hat{y}_i'$ et $V_I(\hat{\mu}_{y_i}) = n^{-2} r^{-1} m \sum_{i=1}^n \hat{e}_i'^2$. Par conséquent,

$$V(\hat{\mu}_{y_i}) = V \left(n^{-1} \sum_{i=1}^n \hat{y}_i' \right) + E \left(n^{-2} r^{-1} m \sum_{i=1}^n \hat{e}_i'^2 \right) \quad (A.2)$$

Maintenant, par un argument similaire à celui qui nous a mené à (2), nous obtenons

$$\text{Var} \left(n^{-1} \sum_{i=1}^n \hat{y}_i' \right) = [n^{-1} R_2 + r^{-1} (1 - R_2)] \sigma_{y_i}^2. \quad (A.3)$$

Puisque $\hat{y}_i' - \bar{y}_i' = (\mathbf{x}_i' - \bar{\mathbf{x}}') \beta + o_p(1)$, nous appliquons la théorie classique de régression pour obtenir

$$E \left[(r - p)^{-1} \sum_{i=1}^r \hat{e}_i'^2 \right] = (1 - R_2) \sigma_{y_i}^2 \quad (A.4)$$

et

$$E \left[(n - 1)^{-1} \sum_{i=1}^n (\hat{y}_i' - \bar{y}_i')^2 \right] = R_2 \sigma_{y_i}^2. \quad (A.5)$$

Par conséquent, l'équation (10) est prouvée et l'estimateur donné par (12) est cohérent pour la variance donnée par l'équation (10).

[M1] Imputation par la méthode hot deck pondérée avec remise considérée par Rao et Shao (1992), où une valeur manquante est remplacée par imputation d'une valeur sélectionnée au hasard parmi les répondants avec remise et probabilité proportionnelle au poids de sondage.

[M2] Imputation par la méthode hot deck pondérée sans remise, qui est identique à la méthode [M1], hormis le fait que la sélection est effectuée par échantillonnage sans remise. La sélection de donneurs sans remise est exécutée systématiquement selon la méthode décrite par Hansen, Hurwitz et Madow (1953, page 343) pour les répondants tirés par ordre aléatoire.

[M3] Imputation globale de la moyenne, où l'on impute la moyenne pondérée calculée pour les répondants qui figurent dans l'échantillon.

Par conséquent, toutes les méthodes d'imputation comportent une cellule d'imputation unique qui regroupe toutes les strates.

Pour chaque ensemble contenant de données imputées, nous calculons trois estimateurs de la variance, à savoir V_n , qui est l'estimateur naïf de la variance obtenu lorsque l'on traite les données imputées comme s'il s'agissait de données observées, V_o , qui est l'estimateur jackknife corrigé de la variance proposé par Rao et Shao (1992) pour [M1] et [M2] et par Rao et Sitter (1995) pour [M3], et V^* , qui est l'estimateur jackknife de la variance fondé sur les pseudo-données. Nous produisons l'ensemble de pseudo-données au moyen de l'équation (24) pour [M3]. L'estimateur complet de la variance d'échantillon est un estimateur jackknife type pour l'échantillonnage par grappe stratifié, selon lequel une grappe est supprimée à chaque itération. Notons que l'estimateur jackknife type est un estimateur cohérent de la variance dans les conditions du modèle avec correction intra-grappe non nulle. Donc, nous pouvons appliquer la méthode type du jackknife fondée sur les pseudo-données à l'ensemble de données considéré. Les estimateurs ponctuels de la moyenne de population sont non biaisés dans le cas des trois scénarios distincts d'imputation et nous ne les présentons pas ici.

Le tableau 2 donne le biais relatif des estimateurs de la variance, l'écart-type du biais relatif des estimateurs de la variance et le coefficient de corrélation d'échantillon entre l'estimateur jackknife corrigé de la variance de Rao et le nouvel estimateur de la variance fondés sur les 5 000 échantillons. Nous calculons le biais relatif de V^* , en tant qu'estimateur de la variance de V^* , au moyen de $[\text{Var}_B(V^*)]^{-1} [E_B(V^*) - \text{Var}_B(V^*)]$, où l'indice B représente la distribution produite par la simulation de Monte Carlo. Enfin, nous calculons la variance de corrélation des deux estimateurs de la variance pour donner une mesure de la linéarité relative de ces estimateurs.

Tableau 2
Biais relatif de l'estimateur de la variance, écart-type du biais relatif et coefficient de corrélation de l'échantillon entre l'estimateur de la variance de Rao et le nouvel estimateur de la variance fondés sur 5 000 échantillons

Taux de réponse (p)	Méthode d'imputation	Biais rel. $\times 100$ (E. - T. $\times 100$)	Rao	Nouveau de corr.
0,9	M1	-17,40 (2,02)	1,61 (2,03)	1,70 (2,04)
	M2	-17,50 (2,00)	1,41 (2,01)	0,81 (2,03)
	M3	-18,03 (2,03)	1,16 (2,05)	1,15 (2,04)
0,8	M1	-34,45 (2,01)	0,65 (2,03)	0,49 (2,05)
	M2	-32,89 (2,01)	2,49 (2,04)	0,19 (2,03)
	M3	-34,96 (2,01)	1,59 (2,03)	1,59 (2,03)
0,7	M1	-48,96 (2,01)	0,21 (1,99)	0,41 (2,04)
	M2	-44,76 (2,02)	5,31 (2,05)	0,76 (2,05)
	M3	-50,21 (2,02)	1,53 (2,05)	1,52 (2,04)
0,6	M1	-59,80 (2,02)	1,58 (2,05)	1,27 (2,06)
	M2	-54,86 (2,03)	7,10 (2,07)	-0,75 (2,07)
	M3	-64,11 (2,00)	-0,35 (2,04)	-0,35 (2,01)
0,5	M1	-69,75 (1,99)	0,84 (2,03)	1,12 (2,03)
	M2	-59,90 (2,01)	15,07 (2,07)	2,27 (2,06)
	M3	-74,44 (1,97)	1,99 (2,00)	1,98 (2,00)

Le tableau 2 confirme notre théorie de la façon suivante.

1. Il est bien connu que l'estimateur naïf de la variance sous-estime fortement la variance réelle. L'estimateur jackknife corrigé de la variance donne de bons résultats pour [M1] et [M3], mais pas pour [M2]. La théorie qui sous-tend la méthode corrigée du jackknife suppose que les imputations par la méthode hot deck s'appuient sur l'échantillonnage avec remise, lequel n'est pas utilisé dans [M2]. À mesure que le taux de réponse diminue dans le tableau 2, le biais relatif de l'estimateur jackknife corrigé augmente.

2. La nouvelle méthode basée sur les pseudo-données donne de bons résultats même pour l'imputation avec échantillonnage sans remise [M2]. Comme nous en avons discuté à la fin de la section 3, la formule (29) suffit, à elle seule, à produire les pseudo-données pour une grande classe de méthodes d'imputation.

3. Comme le montrent les coefficients de corrélation, les comportements de l'estimateur jackknife corrigé et de l'estimateur proposé de la variance sont très semblables pour l'imputation de la moyenne [M3], parce que les deux estimateurs sont asymptotiquement équivalents, comme nous l'avons exposé à la section 5.

7. CONCLUSIONS

Nous avons décrit certaines méthodes qui permettent de produire des pseudo-données pour estimer la variance.

Pour éprouver la théorie qui précède, nous avons réalisé une étude en simulation portant sur une population artificielle, finie, à partir de laquelle nous avons sélectionné des échantillons répétés. La population compte $L = 32$ strates, h grappes dans la strate h et 20 unités finales dans chaque grappe. Les valeurs des paramètres de population ont été choisies de façon à ce qu'elles correspondent aux valeurs de population réelles observées dans le cadre de la U.S. (National Assessment of Educational Progress Study (Hansen et Tepping 1985) et sont énumérées au tableau 1.

Nous considérons un plan d'échantillonnage par grappe stratifié où on sélectionne $n_h = 2$ grappes avec remise dans la strate h avec probabilité égale et où toutes les unités finales qui figurent dans les grappes sélectionnées sont dans l'échantillon. La fraction d'échantillonnage est de 6,4 %. Pour chaque unité échantillonnée y_{hi} , nous produisons une variable indicatrice de réponse a_{hi} à partir de

et cet a_{ij} est indépendant de y_{ij} . Les valeurs de p considérées pour la simulation sont $p = 0,9, 0,8, 0,7, 0,6$ et $0,5$. Nous sélectionnons un ensemble de 5 000 échantillons selon le même plan d'échantillonnage. Pour chaque échantillon sélectionné, nous considérons trois méthodes d'imputation.

décrits pour l'imputation déterministe à la sous-section qui précède. En premier lieu, définissons la fonction indicatrice

$$d_{ij} = \begin{cases} 1 & \text{si l'unité } i \text{ est utilisée comme} \\ & \text{donneur pour l'unité } j \\ 0 & \text{autrement.} \end{cases} \quad (27)$$

Ainsi, l'estimateur de la moyenne de y en recourant à l'imputation aléatoire est

$$\bar{y}_I = \sum_{i=1}^I w_i y_i^* \quad (28)$$

où

$$y_i^* = y_i + a_i(1 + d_i)(y_i - y_j) \quad (29)$$

et

$$d_i = \sum_{j=1}^J (1 - a_j) d_{ij} w_i^{-1} w_j. \quad (30)$$

Si les poids d'échantillonnage originaux sont les mêmes, alors d_i représente le nombre de fois que l'unité i est utilisée comme donneur. Nous supposons que

$$E[a_i(1 + d_i) | F_1] = 1 \quad (31)$$

où $F_1 = \{(i, \mathbf{x}_i, y_i); i = 1, 2, \dots, n\}$. Dans (31), l'espérance a trait à la distribution combinée du mécanisme de réponse et du mécanisme d'imputation. Alors, nous avons

$$E(\bar{y}_I | F_1) = \bar{y}.$$

Si nous supposons que les probabilités de réponse sont égales, alors, en vertu de l'hypothèse (31), la probabilité de sélection des donneurs devrait être proportionnelle aux poids, ce qui correspond aux conditions établies par Rao et Shao (1992) pour l'imputation aléatoire.

Maintenant, supposons que

$$\bar{y}_I^n = \sum_{i=1}^I w_i [y_i^* + a_i(1 + d_i)(y_i - y_j)] \quad (32)$$

où $y_i^* = \mathbf{x}_i' \beta$. Alors, nous avons aussi $\bar{y}_I = \bar{y}_I^n + (\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1)'(\beta - \beta) = \sum_{i=1}^I w_i a_i(1 + d_i) \bar{\mathbf{x}}_i$. En vertu de l'hypothèse (31), nous avons $E(\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 | F_1) = 0$. Si les contrastes sont peu sévères, $\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 = O_p(n^{-1/4})$ et $\bar{y}_I = \bar{y}_I^n + O_p(n^{-1})$. Maintenant,

$$V(\bar{y}_I^n - \bar{y}_I^n) = V[E(\bar{y}_I^n | \mathbf{a}, \mathbf{d})] + E[V(\bar{y}_I^n - \bar{y}_I^n | \mathbf{a}, \mathbf{d})] \quad (33)$$

où $\mathbf{d} = (d_1, d_2, \dots, d_N)$. Subordonné à \mathbf{a} et à \mathbf{d} , l'estimateur \bar{y}_I^n est linéaire. Donc, nous pouvons utiliser les pseudo-données

pour estimer la variance de \bar{y}_I .

5. COMPARAISON À LA MÉTHODE CORRIGÉE DU JACKKNIFE

Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance pour résoudre le problème de l'imputation par la méthode du quotient. Dans les conditions décrites à la section 4, l'estimateur imputé par quotient de μ_y est

$$\mu_I = \sum_{i=1}^I w_i [a_i y_i + (1 - a_i) y_j] \quad (34)$$

où $y_i^* = x_i' \hat{R}$ et $\hat{R} = (\sum_{i=1}^I w_i a_i x_i)^{-1} \sum_{i=1}^I w_i a_i y_i$. L'estimateur de la variance de Rao et Sitter (1995) est donné par

$$V_a = \sum_{i=1}^I c_i (\mu_I^{(i)} - \mu_I)^2, \quad (35)$$

où la répétition jackknife corrigée obtenue à la i^{e} itération est donnée par

$$y_i^{*(*)} = \begin{cases} x_i' \hat{R}^{(i)} & \text{si } a_i = 1 \\ x_i' \hat{R} & \text{si } a_i = 0 \end{cases} \quad (36)$$

avec $\hat{R}^{(i)} = (\sum_{j=1}^I w_j M_{(i)}^j a_j x_j)^{-1} \sum_{j=1}^I w_j M_{(i)}^j a_j y_j$. Les valeurs corrigées (36) de la méthode de Rao et Sitter (1995) peuvent aussi être considérées comme des pseudo-données pour l'estimation de la variance. Notons que le calcul des pseudo-données (36) oblige à recalculer $\hat{R}^{(i)}$ pour chaque i , avec $a_i = 1$.

Nous modifions le calcul des pseudo-valeurs y_i^* dans (5) pour obtenir

$$y_i^* = \begin{cases} y_i & \text{si } a_i = 0 \\ y_i + c_i \left(\frac{x_i}{\bar{x}_1} \right) (y_i - y_j) & \text{si } a_i = 1, \end{cases} \quad (37)$$

où $\bar{x}_2 = \sum_{i=1}^I w_i r_i^{-1} n a_i x_i$, $\bar{x}_1 = n^{-1} \sum_{i=1}^I w_i x_i$ et $c_i = r_i^{-1} n$. Nous insérons le terme (x_i / \bar{x}_2) pour améliorer les propriétés conditionnelles de y_i^* , étant donné l'échantillon de première phase. L'estimateur résultant de la variance est approximativement équivalent à l'estimateur jackknife corrigé de la variance donné par (34). Pour le confirmer, notons que les valeurs corrigées (35) peuvent s'écrire sous la forme

$$\mu_I^{(i)} = \frac{\sum_{j=1}^I w_j M_{(i)}^j a_j y_j}{\sum_{j=1}^I w_j M_{(i)}^j a_j x_j} =: Z^{(i)} S^{(i)},$$

Ici, w_i^* représente le poids d'échantillonnage de l'unité i dans l'échantillon de deuxième phase et est défini par

$$w_i^* = \left[\text{Tr}(\text{ soit dans l'échantillon de deuxième phase } i \text{ soit dans l'échantillon de première phase}) \right]^{-1} w_i.$$

En outre, $\sum_{i=1}^N w_i^* = 1$. Si nous supposons que l'échantillon de deuxième phase est un échantillon aléatoire de taille r sélectionné à partir de l'échantillon de première phase n , alors $w_i^* = nr^{-1} w_i$. Dans certaines conditions, nous pouvons écrire l'estimateur donné par (17) sous la forme

$$(18) \quad \bar{y}_I = \sum_{i=1}^I w_i y_i^*.$$

La représentation (18) est vérifiée si $(w_i^*)^{-1} w_i$ se trouve dans l'espace colonne de la matrice $\mathbf{X} = (\mathbf{x}_1^*, \dots, \mathbf{x}_I^*)'$ parce qu'alors, nous avons $\sum_{i=1}^I w_i (\mathbf{y}_I - \mathbf{y}_i^*) = 0$ provenant de $\sum_{i=1}^I w_i^* \mathbf{x}_i' (\mathbf{y}_I - \mathbf{y}_i^*) = 0$. Nous supposons une série d'échantillons et de populations finies telle que celle décrite dans Fuller (1998). Définissons $\bar{\mathbf{x}}_1 = \sum_{i=1}^n w_i \mathbf{x}_i$ et $(\bar{\mathbf{x}}_2, \bar{y}_2) = \sum_{i=1}^I w_i^* (\mathbf{x}_i, y_i)$. Nous adoptons aussi les mêmes hypothèses que dans Fuller (1998). C'est-à-dire

$$(19) \quad E(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2) = (\mu_x, \mu_x, \mu_y),$$

et

$$(20) \quad V\left\{ \left(\bar{\beta} - \beta \right), \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2 \right\} = O(n^{-1}),$$

où $(\mu_x, \mu_y) = N^{-1} \sum_{i=1}^N (\mathbf{x}_i, y_i)$ et $\beta = \left(\sum_{i=1}^I \mathbf{x}_i' \mathbf{x}_i' \right)^{-1} \sum_{i=1}^I \mathbf{x}_i' y_i$. Pour $i = 1, 2, \dots, N$, définissons

$$a_i = \begin{cases} 1 & \text{si l'unité } i \text{ répond à l'échantillonnage} \\ 0 & \text{autrement,} \end{cases}$$

et $\mathbf{a} = (a_1, a_2, \dots, a_N)$. La définition étendue de a_i est dictée par Fay (1991) et utilisée par Shao et Steel (1999). Maintenant, supposons que

$$(21) \quad \bar{y}_I = \sum_{i=1}^I w_i y_i^*$$

où

$$(22) \quad y_i^* = y_i + a_i w_i^{-1} w_i^* (y_i - y_i^*)$$

avec $\bar{y}_I = \bar{\mathbf{x}}_1 \beta$. Alors, nous avons $\bar{y}_I = \bar{y}_I + O_p(n^{-1})$ et $\bar{y}_I = \bar{y}_I + O_p(n^{-1})$. Maintenant, $V(\bar{y}_I - \bar{y}_I) = o_p(n^{-1})$. Maintenant,

$$(23) \quad V(\bar{y}_I - \bar{y}_I) = V[E(\bar{y}_I - \bar{y}_I | \mathbf{a})] + E[V(\bar{y}_I - \bar{y}_I | \mathbf{a})].$$

Le premier terme du deuxième membre de (23) est nul, car $E(\bar{y}_I - \bar{y}_I | \mathbf{a}) = 0$ en vertu du modèle (7). Pour estimer le deuxième terme de (23), notons que, subordonné à \mathbf{a} , \bar{y}_I est un estimateur linéaire. Donc, la méthode standard

d'estimation de la variance appliquée à l'ensemble de pseudo-données $\tilde{\mathbf{Y}}^* = \{y_i^*, i = 1, 2, \dots, n\}$ produira une estimation non biaisée de la variance de $\bar{y}_I = \sum_{i=1}^n w_i y_i^*$. Puisque l'ensemble $\tilde{\mathbf{Y}}^*$ ne peut être observé, nous pouvons utiliser l'ensemble $\mathbf{Y}^* = \{y_i^*, i = 1, 2, \dots, n\}$, où

$$(24) \quad y_i^* = y_i + a_i w_i^{-1} w_i^* (y_i - y_i^*)$$

Pour montrer que nous pouvons utiliser l'ensemble \mathbf{Y}^* pour obtenir un estimateur de la variance cohérent, supposons que l'estimateur de la variance de l'échantillon complet de \bar{y} peut être représenté par

$$V = \sum_{i=1}^I c_i (\bar{y}^{(i)} - \bar{y})^2$$

où L est le nombre de répétitions, c_i est le i^{e} facteur de répétition et $\bar{y}^{(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j$ est la i^{e} répétition de \bar{y} . Le terme $M_j^{(i)}$ est le multiplicateur de répétition appliqué au poids de l'unité j lors de la i^{e} répétition. Par exemple, en cas d'échantillonnage aléatoire simple, le multiplicateur jackknife est

$$M_j^{(i)} = \begin{cases} 0 & \text{si } i = j \\ (n - 1)^{-1} n & \text{si } i \neq j \end{cases}$$

Supposons que nous appliquons l'estimateur répété de la variance V à l'ensemble \mathbf{Y}^* pour obtenir

$$V^* = \sum_{i=1}^I c_i (\bar{y}^{*(i)} - \bar{y}_I)^2$$

où $\bar{y}^{*(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j^*$ avec y_j^* défini dans (24). Alors, nous obtenons

$$(25) \quad \bar{y}^{*(i)} - \bar{y}_I = \bar{y}^{(i)} - \bar{y}_I + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) (\beta - \beta)$$

où

$$(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \sum_{j=1}^n w_j M_j^{(i)} (\mathbf{x}_j, a_j w_j^{-1} w_j^* \mathbf{x}_j).$$

Nous montrons à l'annexe C que,

$$(26) \quad V^* = \sum_{i=1}^I c_i \left(\bar{y}^{(i)} - \bar{y}_I \right)^2 + o_p(n^{-1}).$$

Par conséquent, nous pouvons utiliser l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données \mathbf{Y}^* comme approximation de l'estimateur jackknife type de la variance appliqué à l'ensemble de pseudo-données \mathbf{Y}^* .

4.2 Imputation aléatoire

Les arguments concernant l'estimation de la variance en cas d'imputation aléatoire sont fort semblables à ceux

Supposons que le modèle est estimé et que les valeurs imputées sont données par

$$y_i = y_i' + \varepsilon_i, \quad i = r+1, r+2, \dots, n \quad (8)$$

où $y_i' = \mathbf{x}_i' \hat{\beta}$ avec $\hat{\beta} = (\sum_{i=1}^r \mathbf{x}_i' \mathbf{x}_i)^{-1} \sum_{i=1}^r \mathbf{x}_i' y_i'$ et où ε_i est choisi au hasard dans l'ensemble $\hat{\varepsilon}_r = \{\varepsilon_i = y_i - y_i'; i = 1, 2, \dots, r\}$. L'estimateur de la moyenne y est

$$\bar{y}_y = n^{-1} \sum_{i=1}^r y_i' \quad (9)$$

où $y_i' = y_i$, si $i = 1, 2, \dots, r$.

Si nous choisissons les ε_i avec remise et avec probabilité

égale à partir de l'ensemble $\hat{\varepsilon}_r$, alors la variance $\hat{\mu}_y$ est donnée, approximativement, par

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}m)(1 - R^2)]\sigma_y^2 \quad (10)$$

où $m = n - r$ et où R^2 représente le carré du coefficient de corrélation multiple entre y et \mathbf{x} . L'augmentation de la variance due à l'utilisation de l'imputation aléatoire avec ε_i , plutôt que l'utilisation de $\varepsilon_i \equiv 0$, est $n^{-2}m(1 - R^2)\sigma_y^2$. Par conséquent, un estimateur de la variance de la moyenne d'échantillon imputée est donné par (6) où le c_p de (4) est donné par

$$c_I = [n(n-1)(r^{-1} + n^{-2}m)(r - p)^{-1}]^{1/2}, \quad (11)$$

et p représente la dimension de β . Nous avons

$$V\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^r (\hat{y}_i - \bar{y}_y)^2$$

$$+ (r^{-1} + n^{-2}m)(r - p)^{-1} \sum_{i=1}^r (y_i - \hat{y}_i)^2 \quad (12)$$

Supposons que dans l'imputation.

nombre de fois qu'une valeur \hat{e} est utilisée comme donneur que la valeur y réponde. En outre, nous enregistrons le sélection sont inversement proportionnelles à la probabilité aucune contrainte, si ce n'est que les probabilités de sélection aléatoire pour l'imputation, mais n'imposons variance, nous supposons que l'on applique une méthode de

Pour considérer une autre méthode d'estimation de la imputées. Les calculs de (10) et (12) figurent à l'annexe A. tous les échantillons possibles contenant des données variance non conditionnelle, c'est-à-dire la moyenne sur utilisant c_I de l'équation (11) est un estimateur de la où $\bar{y}_I = \sum_{i=1}^n y_i'$. L'estimateur de la variance calculé en

et que

$$y_i' = \begin{cases} y_i + c_i(y_i - \hat{y}_i) & i = 1, 2, \dots, r \\ \hat{y}_i & i = r+1, r+2, \dots, n \end{cases} \quad (13)$$

où

$$\bar{y}_I = \sum_{i=1}^r w_i' y_i' + \sum_{i=r+1}^n w_i' \hat{y}_i \quad (17)$$

de y dans des conditions d'imputation par régression est

Si les r premiers éléments sont observés et que les $n - r$ autres éléments manquent, alors l'estimateur de la moyenne

complet de la moyenne de y peut être représenté par simple. Supposons que l'estimateur pour l'échantillon sondage complexes, ainsi qu'à l'échantillonnage aléatoire La méthode proposée est applicable à des plans de

4.1 Imputation déterministe

4. PLANS DE SONDAGE COMPLEXES

applicabilité est très grande.

où nous utilisons la notation I pour représenter l'espérance.

$$E_I\{1 + d_I y^2\} = \left(\frac{r}{n}\right)^2 + \frac{r}{m} \left(1 - \frac{r}{n}\right) \quad (16)$$

à l'échantillon et aux répondants,

Si la méthode d'imputation est une méthode d'échantillonnage aléatoire simple avec remise, alors, subordonnée

$$+ n^{-2}r(r-p)^{-1} \sum_{i=1}^r (1 + d_I y_i^2)(y_i - \hat{y}_i)^2. \quad (15)$$

représente par

$$c_p = [n^{-1}(n-1)r(r-p)^{-1}]^{1/2}(1 + d_I) \quad (14)$$

variance à chaque ensemble de données pour calculer le terme de variance à l'intérieur de l'ensemble de données, tandis que la méthode de Rao et celle de Särndal consiste à appliquer un estimateur particulier de la variance à l'ensemble contenant des données imputées. Ici, nous proposons une méthode pour créer un ensemble de pseudo-données unique pour l'estimation de la variance. À la section 2, nous présentons la nouvelle méthode dans le cadre de l'échantillonnage à deux phases. À la section 3, nous illustrons l'extension de la méthode proposée à la méthode d'imputation aléatoire. À la section 4, nous étendons la méthode proposée aux plans d'échantillonnage complexes. À la section 5, nous comparons les résultats à ceux de l'estimateur jackknife corrigé de la variance. À la section 6, nous présentons une étude en simulation limitée. Enfin, nous présentons certaines conclusions à la section 7. L'exposé de certaines preuves figure en annexe.

2. UNE MÉTHODE D'ESTIMATION DE LA VARIANCE

Nous décrivons une méthode d'estimation de la variance applicable aux échantillons à deux phases et aux échantillons contenant des données imputées. En plus de l'ensemble de données totales, la méthode nécessite un ensemble distinct de données pour l'estimation de la variance. Pour présenter la méthode et illustrer les concepts, considérons un échantillon à deux phases. Supposons que la deuxième phase consiste en un échantillon aléatoire simple de taille r sélectionné à partir de l'échantillon de première phase, qui est un échantillon aléatoire simple de taille n sélectionné à partir d'une population infinie. Supposons que l'estimateur par régression de la moyenne d'une caractéristique y est

$$\hat{y}_y = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}, \quad (1)$$

où

$$(\bar{y}_2, \bar{x}_2) = r^{-1} \sum_{i=1}^r (y_i, x_i),$$

$$\bar{x}_1 = n^{-1} \sum_{i=1}^n x_i,$$

$$\hat{\beta} = \left[\sum_{i=1}^r (x_i - \bar{x}_2)(x_i - \bar{x}_2)^2 \right]^{-1} \sum_{i=1}^r (x_i - \bar{x}_2)(y_i - \bar{y}_2)$$

et que les unités de deuxième phase sont marquées d'un indice dont la valeur varie de un à r . Il est bien connu (par exemple, Cochran 1977, équation 12.51) que la variance de l'estimateur par régression est, approximativement,

$$V\{\hat{y}_y\} = [n^{-1}p_2 + r^{-1}(1-p_2)]\sigma_y^2, \quad (2)$$

où le premier élément de chaque \mathbf{x}_i est égal à 1 et où les e_i sont des variables aléatoires non corrélées $(0, \sigma_e^2)$.

où p est la corrélation de population entre y et x et où σ_y^2 est la variance de population de y . Selon la théorie classique de la régression, un estimateur de la variance prend la forme

$$V\{\hat{y}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + r^{-1}(r-2)^{-1} \sum_{i=1}^r (\hat{y}_i - \bar{y})^2 \quad (3)$$

où $\hat{y}_i = \bar{y}_2 + (x_i - \bar{x}_2)\hat{\beta}$ pour $i = 1, 2, \dots, n$, et $\bar{y}_i = n^{-1} \sum_{j=1}^n \hat{y}_j$. Notons que \bar{y}_i est une autre façon d'écrire \hat{y}_i dans (1).

Supposons que

$$\hat{y}_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, \dots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i) & i = 1, 2, \dots, r. \end{cases} \quad (5)$$

Alors,

$$V\{\hat{y}_y\} = n^{-1}(n-1)^{-1} \sum_{i=1}^n (y_i^* - \bar{y})^2 \quad (6)$$

où \bar{y}_i représente la moyenne des y_i^* , ainsi que la moyenne des y_i , parce que la somme des termes $y_i - \hat{y}_i$ est nulle. L'équation (6) est la forme opérationnelle de l'estimateur proposé. L'ensemble de données pour l'estimation de la variance contient la pseudo-observation y_i^* .

Dans la mesure où le modèle utilisé pour l'imputation correspond au modèle d'échantillonnage à deux phases, les données manquantes le sont au hasard et que nous utilisons la régression pour imputer les valeurs manquantes de y_i , alors l'équation (6) est applicable immédiatement. Naturellement, l'imputation par régression ou l'échantillonnage à deux phases peuvent comprendre l'utilisation d'un vecteur \mathbf{x} .

3. EXTENSIONS À L'IMPUTATION ALÉATOIRE

Une extension modérée de la méthode décrite à la section 2 nous permet d'estimer la variance d'une moyenne d'échantillon par imputation aléatoire. En fait, d'autres méthodes sont possibles.

Par exemple, supposons que le modèle d'imputation est le modèle de régression

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + e_i \quad (7)$$

Estimation de la variance après imputation

JAE-KWANG KIM¹

RÉSUMÉ

On recourt fréquemment à l'imputation pour compenser la non-réponse partielle. L'estimation de la variance après imputation a suscité de nombreux débats et plusieurs estimateurs ont été proposés. Nous proposons un estimateur de la variance fondé sur un ensemble de pseudo-données créé uniquement pour estimer la variance. L'application des estimateurs type de la variance de données complètes à l'ensemble de pseudo-données produit des estimateurs cohérents dans le cas des estimateurs linéaires pour diverses méthodes d'imputation, y compris l'imputation par la méthode hot deck sans remise et avec remise. Nous illustrons l'équivalence asymptotique de la méthode proposée et de la méthode corrigée du jackknife de Rao et Sitter (1995). La méthode proposée s'applique directement à l'estimation de la variance en cas d'échantillonnage à deux phases.

MOTS CLÉS : Échantillonnage à deux phases; non-réponse partielle; imputation déterministe; imputation aléatoire.

1. INTRODUCTION

L'imputation, c'est-à-dire l'insertion de valeurs pour remplacer les réponses manquantes, est utilisée fréquemment lorsque des données d'enquête manquent. L'un des avantages de l'imputation est sa commodité. Autrement dit, nous pouvons appliquer aux ensembles contenant des données imputées des programmes types conçus pour des données complètes en vue de calculer des estimations ponctuelles. Rubin (1996), Fay (1996) et Rao (1996) ont

passé en revue divers problèmes posés par l'imputation. Après avoir précisé ce dernier, nous pouvons recourir à l'imputation déterministe ou à l'imputation aléatoire selon le modèle. Dans le cas de l'imputation aléatoire, on procède à une forme d'échantillonnage probabiliste en vue d'imputer des données pour remplacer les valeurs manquantes. Nous donnons à ce mécanisme aléatoire supplémentaire le nom de mécanisme d'imputation. L'imputation déterministe, quant à elle, ne nécessite pas l'ajout d'un mécanisme aléatoire supplémentaire. Si nous considérons l'ensemble de répondants comme un échantillon aléatoire de l'échantillon original, nous donnons au mécanisme de sélection des répondants le nom de mécanisme de réponse. Ce dernier est souvent considéré comme étant la deuxième phase d'échantillonnage. Pour plus de précisions, consulter Särndal et Swensson (1987).

Le choix d'un modèle et d'une méthode d'imputation appropriés permet de réduire considérablement le biais dû à la non-réponse qui entache les résultats lorsque l'on se sert uniquement des données observées. Cependant, il est reconnu qu'un estimateur de la variance qui traite les données imputées comme s'il s'agissait de données observées est incohérent. Diverses méthodes ont été proposées pour estimer la variance après l'imputation. Rubin et Schenker (1986) et

Rubin (1987) recommandent l'imputation multiple qui crée plusieurs ensembles de données et produit des statistiques complètes pour chaque ensemble contenant des données imputées. L'estimateur de la variance est calculé par combinaison de deux termes, à savoir la variance dans les ensembles de données et la variance entre les ensembles de données. La méthode d'imputation multiple consiste à appliquer un estimateur type de la variance à chaque ensemble de données pour calculer les termes de variance à l'intérieur de l'ensemble de données et des estimateurs ponctuels type pour calculer la variance entre ensembles contenant des données imputées. L'application de cette méthode dépend du choix d'une méthode d'imputation appropriée. Autrement dit, l'imputation devrait remplir les conditions 1 à 3 décrites par Rubin (1987, pages 118 et 119). Ces conditions ne sont pas toujours faciles à satisfaire. (Par exemple, voir Fay 1992.) Même dans Schaffer (1997), il n'est pas montré que les méthodes d'imputation multiples décrites sont appropriées au sens de Rubin. Comme le fait remarquer Rao (1996), certaines méthodes d'imputation utilisées couramment, y compris l'imputation par la méthode hot deck et l'imputation par régression, ne conviennent pas.

Rao et Shao (1992) et Rao et Sitter (1995) ont proposé un estimateur jackknife corrigé de la variance. La méthode proposée est applicable à plusieurs méthodes d'imputation et à plusieurs plans d'échantillonnage. Le calcul réel, au moyen d'un logiciel standard applicable à des données complètes n'est pas facile, car des calculs spéciaux sont exécutés pour corriger les valeurs imputées pour chaque pseudo-répétition. En outre, Särndal (1992) a proposé une méthode d'estimation de la variance qui tient compte explicitement du modèle utilisé pour l'imputation. Essentiellement, la méthode de Rubin produit plusieurs ensembles de pseudo-données pour l'estimation de la variance et consiste à appliquer un estimateur type de la

¹ Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

ANNEXE 3 : Application de l'estimation composite par régression aux données de l'EPA Illustration au moyen de la méthode axée sur la variation

Matrice X originale

Indicateurs de groupe âge-sexe		Indicateurs de région				
X_1	X_2	X_3	X_4	X_{k+1}	X_p	
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0

Matrice X modifiée pour l'estimation composite quand des $x_i^{(c)}$ sont ajoutées

Indicateurs de groupe âge-sexe		Indicateurs de région					E est l'estimation de l'emploi le mois précédent	
X_1	X_2	X_3	X_4	X_{k+1}	X_p		E'	S'
0	0	1	0	0	0	0	a	0
0	0	1	0	0	0	0	c	0
0	1	0	0	0	0	0	d	0
0	0	0	0	0	0	0	b	0

Taux de
population de
contrôle



Pour les nouvelles unités, fixer a, b, c, \dots de façon à indiquer la situation durant le mois courant (par exemple, $a = 1$ si la personne est occupée, 0 autrement). Pour les unités *appartées*, procéder comme suit :

$$a = e_i + (e_{i-1} - e_i) \times 6/5 \text{ où } e = 1 \text{ si la personne est occupée, } e = 0 \text{ autrement.}$$

$$d = a_i + (a_{i-1} - a_i) \times 6/5 \text{ où } a = 1 \text{ si la personne est occupée en agriculture, } a = 0 \text{ autrement.}$$

Exemples :

- i) Supposons que la personne 2 était employée en agriculture le mois précédent et le mois courant. Alors, $e_{i-1} = e_i = 1$ et $ag_{i-1} = ag_i = 1$; par conséquent, $c = 1 - 0 = 1$ et $d = 1 - 0 = 1$.
- ii) Supposons que la personne 2 était employée en agriculture le mois précédent et qu'elle l'est dans le secteur minier le mois courant. Alors, $e_{i-1} = e_i = 1$, $ag_{i-1} = 1$ et $ag_i = 0$; donc, $c = 1 - 0 = 1$ et $d = (1 - 0) \times 6/5 = 1,2$.

- iii) Supposons que la personne 2 était employée dans le secteur minier le mois précédent et qu'elle est employée en agriculture le mois courant. Alors, $e_{i-1} = e_i = 1$, $ag_{i-1} = 0$ et $ag_i = 1$; donc, $c = 1 - 0 = 1$ et $d = 1 + (0 - 1) \times 6/5 = -0,2$.

BIBLIOGRAPHIE

- BAILLAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.
- CANTWELL, P.J., et ERNST, L.R. (1992). Nouveaux développements dans l'estimation composite pour l'enquête « Current Population Survey ». *Recueil : Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada, 139-149.
- FULLER, W.A., et RAO, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.
- GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. et LINDEYER, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Statistique Canada, numéro de catalogue, 71-526.
- KUMAR, S., et LEE, H. (1983). Evaluation de l'application d'estimateurs composites à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.
- SINGH, A.C., KENNEDY, B. et WU, S. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel. *Techniques d'enquête*, 27, 35-48.
- SINGH, A.C., KENNEDY, B., WU, S. et BRISEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

permettant d'utiliser des valeurs de α différentes pour des variables de contrôle différentes, en continuant d'utiliser un poids final unique pour chaque enregistré.

REMERCIEMENTS

Nous aimerions remercier Avi Singh et le Comité consultatif des méthodes statistiques de Statistique Canada pour leurs contributions à ce projet. Nous sommes aussi redevables à plusieurs personnes dont les commentaires sur des versions antérieures, ont permis de grandement améliorer cet article.

ANNEXE 1

Relation entre α , ρ et (A, K) . Kumar et Lee (1983) ont déterminé les valeurs optimales de A et K de l'estimateur composite AK pour les estimations du niveau et de la variation en fonction du coefficient de corrélation ρ . Nous avons dérivé une relation approximative entre les valeurs de A et K , ρ et α . Puis, nous avons utilisé cette relation pour déterminer de bonnes valeurs de α pour plusieurs variables. Ces valeurs sont présentées dans les tableaux 1 et 2 pour les estimations du niveau et de la variation, respectivement. Les

ANNEXE 2 :

Mesures de la désaisonnalisation pour l'emploi selon la branche d'activité en Ontario

	Valeur F			M7			LISSAGE		
Branche d'activité	greg	$\alpha = 0,60$	$\alpha = 0,75$	greg	$\alpha = 0,60$	$\alpha = 0,75$	greg	$\alpha = 0,60$	$\alpha = 0,75$
Agriculture	87,76	120,18	112,7	0,27	0,23	0,24	37,94	45,36	26,78
Forêtierie	21,34	24,58	23,22	0,5	0,52	0,57	21,76	15,52	15,52
Services publics	4,29	3,48	6,8	1,1	1,25	0,82	15,39	57,5	31,94
Construction	128,3	275,06	246,93	0,26	0,16	0,17	41,68	34,92	23,33
Fabrication	38,22	55,6	69,21	0,37	0,3	0,3	29,02	19,67	19,52
Commerce	9,93	15,12	20,35	0,8	0,68	0,53	25,13	16,2	22,17
Transport	9,16	8,64	9,69	0,94	0,75	0,7	15,36	66,47	19,92
Finances	6,49	8,94	8,84	1,22	0,76	0,77	13,45	44,29	33,46
Serv. Professionnels	5,3	12,91	9,81	1,03	0,72	0,76	12,45	26,27	26,27
Gestion	14,72	24,98	20,35	0,67	0,52	0,52	16,2	44,29	33,46
Education	67,37	219,62	214,37	0,33	0,16	0,19	53,25	44,29	33,46
Services de santé	8,78	10,73	8,48	0,8	0,66	0,75	16,09	44,29	33,46
Information	21,13	52,31	62,94	0,66	0,36	0,35	24,29	44,29	33,46
Hébergement	44,85	75,37	78,03	0,36	0,34	0,3	31,89	44,29	33,46
Autres services	2,61	13,17	12	1,41	0,75	0,81	18,58	44,29	33,46

Valeur F : Valeur F du test exécuté dans le cadre du programme X11-ARIMA pour déceler l'existence d'une saisonnalité stable. L'existence d'une saisonnalité stable est d'autant plus significative que la valeur F est élevée.

M7 : Mesure qui combine les tests pour les saisonnalités stable et mobile. En général, si M7 est supérieur à 1, la série ne présente aucune saisonnalité décelable; par conséquent, elle ne doit pas être rajustée.

LISSAGE : Écart en pourcentage des variations mensuelles dans la série originale et l'écart-type des variations mensuelles dans la série désaisonnalisée. Le lissage que produit la méthode de désaisonnalisation est d'autant plus important que la valeur de l'écart est grande.

- i) retient autant de réponses valides que possible, c'est-à-dire rejeter l'option de supprimer une unité du processus d'estimation;
- ii) élaborer une méthode d'imputation qui ne sous-estime pas l'estimation de la variation de façon significative.

Dans le cas de la non-réponse, on distingue deux situations : le cas A, où un ménage a répondu le mois précédent, mais non le mois courant et le cas B, où la situation est inverse. Dans ce qui suit, i représente une personne dans un ménage affecté.

Cas A : Remplacer y_{it}^n par y_{it}^n . La substitution peut se faire de plusieurs façons. Une méthode simple consiste à remplacer y_{it}^n par la réponse correspondante donnée le mois précédent, c'est-à-dire y_{it-1}^n . Nous avons utilisé cette méthode aux premières étapes de l'étude, mais l'avons rejetée plus tard car elle peut biaiser (sous-estimer) l'estimation de la variation de façon significative. Pour le système d'estimation de l'EPA, nous avons décidé d'utiliser les caractéristiques démographiques et d'emploi connues de la personne pour le mois précédent pour créer des catégories d'imputation puis nous avons appliqué la méthode d'imputation hot deck (c'est-à-dire, données du mois courant) pour obtenir y_{it}^n . Une autre solution consisterait à utiliser une certaine sorte de moyenne.

Cas B : La procédure est analogue; autrement dit, si la valeur du mois précédent manque, on forme des catégories d'imputation en se servant des données recueillies pour le mois t et on se sert de données provenant des unités répondantes au mois $t-1$ pour trouver l'enregistrement donneur.

Si l'unité i a déménagé ou que sa situation par rapport au champ d'observation a été modifiée, les cas qui suivent peuvent se présenter.

Cas C : Supposons que l'unité i était hors du champ d'observation à la période $t-1$, mais dans le champ à la période t (par exemple, une personne qui vient tout juste d'avoir 15 ans ou un immigrant qui vient d'arriver). Alors, la contribution de l'unité i devrait être 0 à la période $t-1$ et y_{it}^n à la période t . Donc, nous posons $x_{it}^n = 0$ puisque $\sum w_{it}^n x_{it}^n$ devrait estimer Y_{t-1}^n .

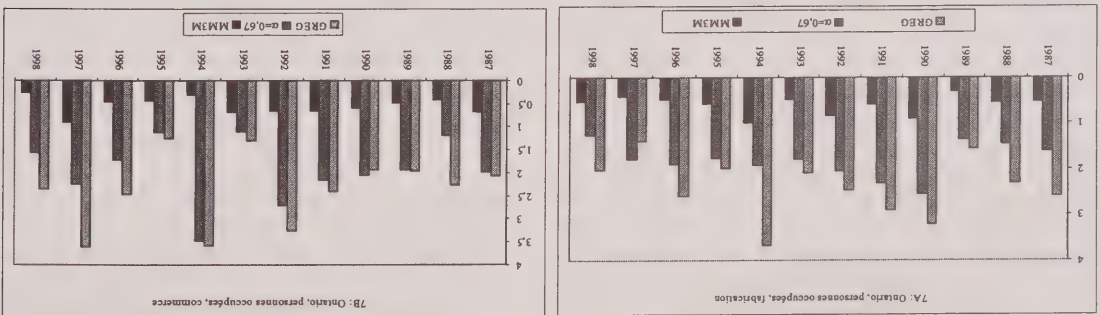
Cas D : Inversement, supposons que l'unité i était dans le champ d'observation le mois précédent et est maintenant hors du champ. Ce scénario inclut, par exemple, les personnes qui ont quitté le pays, sont entrées dans l'armée ou sont décédées. Ces unités devraient être abandonnées, puisque la population cible est la population dans le champ d'observation à la période t (et que l'objectif ultime est d'estimer Y_t^n). Puisque nous échantillonons des logements, mais que nous recueillons des données sur les personnes qui habitent ces logements, deux autres situations peuvent se présenter à cause de l'emménagement de nouvelles personnes dans les logements échantillonnés ou du déménagement de personnes hors de ces logements.

5. CONCLUSIONS

L'estimateur composite permet d'atteindre tous les objectifs établis au début du projet et résout dans l'introduction, le produit des estimations du niveau et de la variation plus efficaces que celles obtenues au moyen de l'estimateur de régression ordinaire, tout en satisfaisant aux contraintes opérationnelles et de cohérence des données. Dans l'ensemble, l'utilisation de l'estimateur composite en choisissant la valeur $\alpha = 2/3$ a un effet modéré sur les nombreuses séries chronologiques produites d'après les données de l'Enquête sur la population active. Lorsque l'effet est important, comme dans le cas de la série sur le commerce en Ontario, la nouvelle série a tendance à être plus compréhensible pour les spécialistes du domaine. Ce genre d'amélioration permet en outre d'expliquer plus facilement les estimations fondées sur les données de l'EPA aux utilisateurs des données et aux membres des médias. Les estimations composites présentent d'autres caractéristiques que les utilisateurs trouveront fort intéressantes. Comme l'estimation composite réduit la variance, il est possible de publier des estimations mensuelles dans de nombreux cas où il fallait se limiter à des moyennes mobiles de trois mois auparavant. De surcroît, l'estimation composite permet de désaisonnaliser convenablement un grand nombre de séries.

L'application de l'estimation composite aux données de l'EPA est une première étape importante. Des études en vue d'améliorer le traitement des erreurs non dues à l'échantillonnage sont en cours et les résultats pourront être intégrés dans le système de pondération et d'estimation n°1 très souple. Le système offre le grand avantage d'être très adaptable. Par exemple, comme on peut modifier facilement la valeur de α , on prévoit comparer les résultats pour toute une gamme de valeurs de α pour un nombre important de personnes dans les logements échantillonnés ou du déménagement de personnes hors de ces logements.

Graphique 7. L'indice d'instabilité



L'indice d'instabilité a été calculé pour seize branches d'activité. Nous présentons ici deux graphiques (7A et 7B), pour les branches de la fabrication et du commerce en Ontario, où sont comparés l'estimateur ordinaire et l'estimateur composite mensuel pour $\alpha = 0,67$ (GREG) la moyenne mobile de trois mois MM3M pour l'estimateur ordinaire. Pour la branche de la fabrication, la valeur moyenne de l'indice pour les estimateurs ordinaire, composite et de la moyenne mobile est de 2,4, 1,8 et 0,60, respectivement. Pour le commerce, les valeurs correspondantes sont 2,4, 1,9 et 0,55. Pour toutes les branches d'activité, l'instabilité des estimations composites est habituellement comprise entre celles des estimations mensuelles ordinaires et des estimations de la moyenne mobile de trois mois. À l'occasion, pour certaines années de référence, les estimations composites sont moins instables que les moyennes mobiles de trois mois ou plus instables que les estimations mensuelles ordinaires, mais, en général, leur instabilité est comprise entre celles des estimations mensuelles ordinaires et des moyennes mobiles de trois mois. Nous constatons aussi que, si les estimations mensuelles sont extrêmement instables, la série d'estimations composites a tendance à être plus stable. Les estimations mensuelles par régression ne concurrencent les estimations composites que si l'indice d'instabilité est faible pour les deux méthodes.

Suite à l'introduction de l'estimation composite, le calcul des moyennes mobiles de trois mois a été abandonné et remplacé par celui des estimations mensuelles, plus souhaitables pour les séries sur les branches d'activité.

Estimations de la variance. Pour les variables ajoutées à titre de totaux de contrôle, comme l'emploi selon la branche d'activité, il est possible de réaliser des gains considérables d'efficacité au niveau provincial, l'efficacité étant définie ici par $\text{Var}(\text{greg})/\text{Var}(\text{composite})$. Pour la plupart des branches d'activité, les gains sont habituellement de l'ordre de 10 % à 20 % mais ils peuvent aller jusqu'à 40 %. Un gain d'efficacité de 40 % équivalait, par exemple, à faire baisser un coefficient de variation de 15 % jusqu'à 12,7 % ou un coefficient de variation de 10 % jusqu'à 8,5 %. Pour les estimations des niveaux provinciaux d'emploi et de chômage, nous observons des gains

Dans tous ces cas, à savoir A, B, C et D, l'objectif est de trouver une solution telle que $\sum_{i \in S} w_i x_i$ soit encore un estimateur de Y_{t-1} . Nous énonçons les deux objectifs qui suivent pour le traitement des cas où des données manquent pour l'un ou l'autre mois de l'échantillon commun :

	A	B	-	C	D
Mois t	X XX...	R RR...	RRR...	RRR...	OOO...
Mois t-1	R RR...	XXX...	RRR...	OOO...	RRR...

Par définition, les variables x_t comprennent des données du mois courant et du mois précédent, ce qui pose des difficultés lorsque l'on ne dispose de données que pour l'un des deux mois pour une personne donnée de l'échantillon commun. Cette situation peut être due à la non-réponse lors de l'un des deux mois, ou à un démenagement ou une modification par rapport au champ d'observation dans l'intervalle de deux mois. Les différents cas possibles sont présentés dans le diagramme qui suit, où R représente une réponse, X représente une non-réponse et O représente une unité hors du champ d'observation.

4. TRAITEMENT DES DONNÉES MANQUANTES

Par définition, les variables x_t comprennent des données du mois courant et du mois précédent, ce qui pose des difficultés lorsque l'on ne dispose de données que pour l'un des deux mois pour une personne donnée de l'échantillon commun. Cette situation peut être due à la non-réponse lors de l'un des deux mois, ou à un démenagement ou une modification par rapport au champ d'observation dans l'intervalle de deux mois. Les différents cas possibles sont présentés dans le diagramme qui suit, où R représente une réponse, X représente une non-réponse et O représente une unité hors du champ d'observation.

Pour les variables non contrôlées, l'estimation composite n'a que peu d'effet, voire aucun, sur l'efficacité, à moins que la variable ne soit fortement corrélée à une variable contrôlée. Par exemple, au niveau provincial, on constate une amélioration de l'efficacité des estimations du nombre d'hommes occupés, parce que cette variable est corrélée au nombre total de personnes occupées, qui est une variable contrôlée. Par contre, pour l'estimation de l'emploi selon la région économique infraprovinciale, on ne constate aucun gain ni perte d'efficacité.

Pour les variables non contrôlées, l'estimation composite n'a que peu d'effet, voire aucun, sur l'efficacité, à moins que la variable ne soit fortement corrélée à une variable contrôlée. Par exemple, au niveau provincial, on constate une amélioration de l'efficacité des estimations du nombre d'hommes occupés, parce que cette variable est corrélée au nombre total de personnes occupées, qui est une variable contrôlée. Par contre, pour l'estimation de l'emploi selon la région économique infraprovinciale, on ne constate aucun gain ni perte d'efficacité.

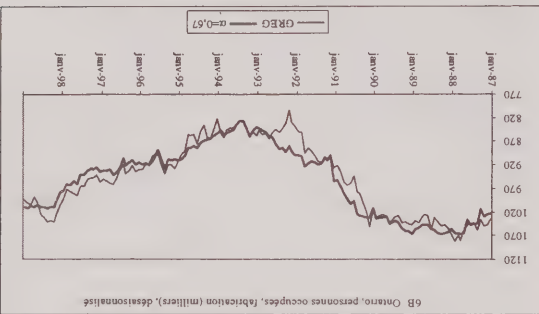
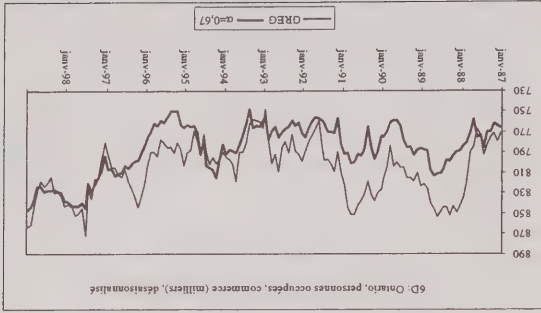
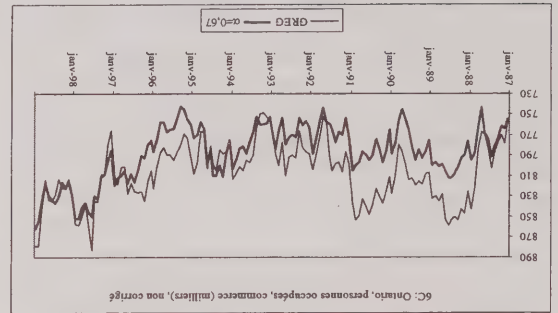
Pour les variables non contrôlées, l'estimation composite n'a que peu d'effet, voire aucun, sur l'efficacité, à moins que la variable ne soit fortement corrélée à une variable contrôlée. Par exemple, au niveau provincial, on constate une amélioration de l'efficacité des estimations du nombre d'hommes occupés, parce que cette variable est corrélée au nombre total de personnes occupées, qui est une variable contrôlée. Par contre, pour l'estimation de l'emploi selon la région économique infraprovinciale, on ne constate aucun gain ni perte d'efficacité.

La comparaison des séries désaisonnalisées (graphique 6D) et non désaisonnalisées (graphique 6C) pour le commerce montre que la désaisonnalisation a un effet assez faible sur la série obtenue par régression, mais modifie considérablement l'illustration de la capacité qu'a l'estimation composite d'acroître suffisamment le rapport signal-bruit pour rendre la désaisonnalisation efficace.

Le programme de désaisonnalisation appliqué aux données de l'Enquête sur la population active calcule diverses mesures qui sont utilisées comme indicateurs de l'efficacité de la désaisonnalisation. Certains de ces mesures, présentées à l'annexe 2, montrent que, pour l'emploi en Ontario durant la période de deux ans allant de 1996 à 1998, le recours à l'estimation composite permet de désaisonnaliser convenablement les données pour un plus grand nombre de branches d'activité. Les résultats pour d'autres provinces et pour le Canada sont similaires.

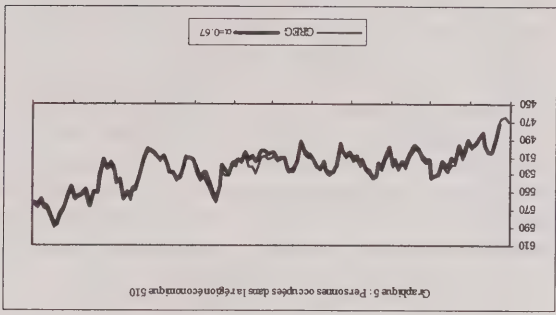
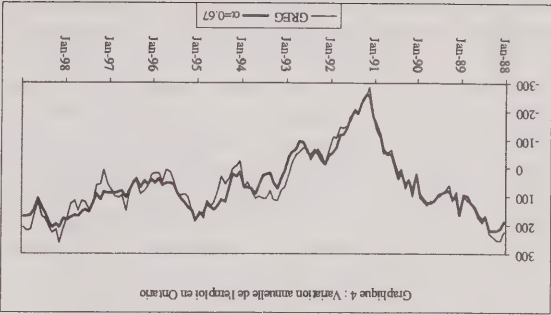
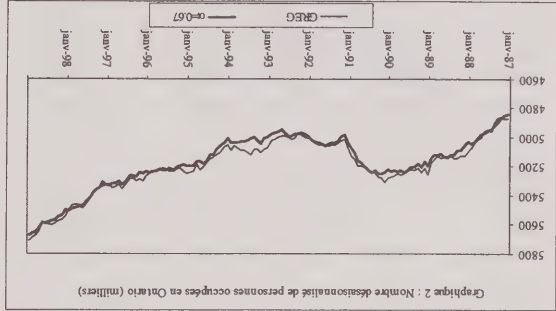
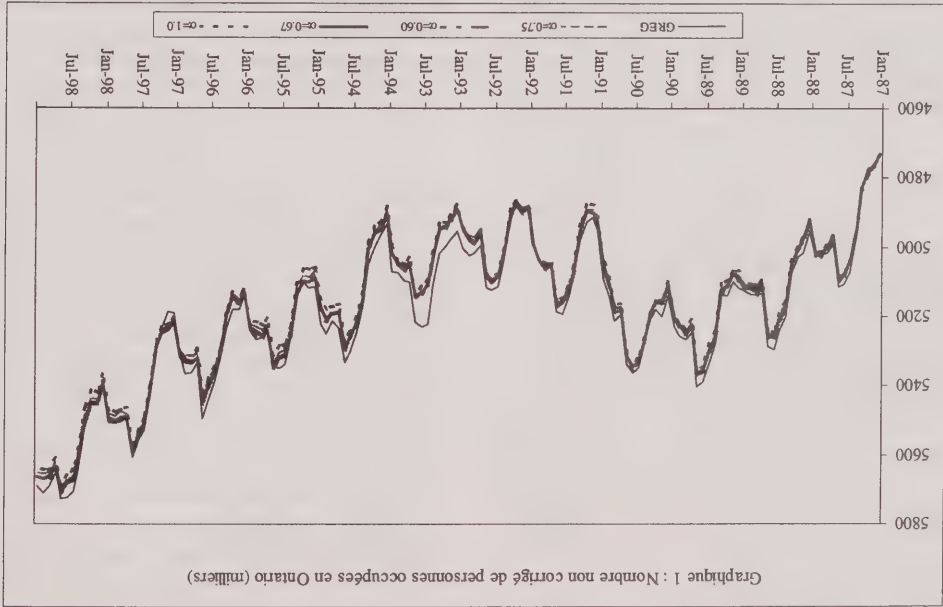
Mesure de la stabilité. Pour plusieurs séries de données importantes, on publiait par le passé les moyennes mobiles de trois ans au lieu des estimations mensuelles, parce que la

Graphique 6. Emploi selon la branche d'activité



forte variabilité d'échantillonnage de ces séries rendait inacceptablement instable la série de données mensuelles. Les estimations provinciales selon la branche d'activité et selon la catégorie professionnelle sont celles auxquelles nous nous intéressons surtout. Nous nous attendions à ce que les estimations composites pour ces séries soient plus stables, donc permettent la publication d'estimations mensuelles au lieu de moyennes mobiles de trois mois. L'indice d'instabilité, qui est une mesure de la stabilité, est calculé comme suit. Pour chaque branche d'activité, on calcule la variation de l'emploi d'un mois à l'autre de ces estimations désaisonnalisées. Puis, on calcule la différence entre les estimations consécutives de la variation. La valeur absolue de cette «variation de la variation» est exprimée en pourcentage de l'estimation mensuelle totale correspondante. Enfin, on calcule la moyenne de ce pourcentage sur l'année complète. Les valeurs de cette mesure sont grandes lorsqu'une série présente de nombreux mouvements consécutifs en directions opposées, ce qui est un signe d'instabilité.

parfois de façon significative des autres. La série d'estimations composées a tendance à être moins instable que celle obtenue par régression. Cette différence est particulièrement évidente pour la série désaisonnalisée sur le commerce que nous avons incluse ici parce qu'elle illustre le cas le plus extrême. Pour cette série, le comportement des estimations par régression obtenues au départ les premières années est difficile à expliquer d'un point de vue technique, qu'il s'agisse de la série désaisonnalisée ou de la série non corrigée. Le comportement de la série sur la fabrication est plus typique des 14 branches d'activité restantes.



maintenant. Autrement dit, l'estimation composite augmente suffisamment le rapport signal-bruit pour rendre la désaisonnalisation efficace. Une conséquence comme de l'estimation composite appréciée par les utilisateurs des données tient au fait que plusieurs estimations publiées antérieurement sous forme de moyenne mobile de trois mois peuvent désormais l'être sous forme d'estimations mensuelles.

Écarts systématiques entre les estimations composites

et ordinaires du niveau. Théoriquement, les espérances, calculées sur tous les échantillons possibles, devraient être les mêmes pour les estimateurs ordinaires et composites, ce qui les rend tous deux non biaisés ou presque non biaisés. On s'attendrait donc à ce que les estimations du niveau obtenues en se servant des deux estimateurs se recoupent au cours du temps. Malheureusement, cela ne se concrétise pas en pratique car, si l'on tient compte des conditions réelles d'enquête, l'estimateur composite et l'estimateur ordinaire n'ont pas la même valeur prévue; consulter, par exemple, Bailar (1975) et Kumar et Lee (1983) pour des résultats sur les estimateurs composites K et AK, respectivement. Kumar et Lee démontrent ce fait en dérivant des expressions explicites pour la valeur prévue de l'estimateur ordinaire et de l'estimateur composite AK. Les échantillons appariés et non appariés diffèrent à cause de différences entre les taux de non-réponse plus élevés et les ménages manquants ont tendance à être de plus petite taille et à avoir un taux d'emploi plus élevé que les ménages répondants. Comme l'estimateur ordinaire et l'estimateur composite produisent des poids différents pour les échantillons appariés et non appariés, leur valeur prévue sera différente. Par conséquent, les séries chronologiques obtenues au moyen des deux estimateurs peuvent présenter des différences systématiques par la variation d'échantillonnage, mais elles sont évidentes pour des séries plus précises, comme celles sur les personnes occupées dans les grandes provinces comme l'Ontario et au Canada. Nos résultats pour les personnes occupées concordent avec ceux décrits par Bailar (1975) pour la *Current Population Survey* aux États-Unis; autrement dit, pour les personnes occupées, les estimations composites ont tendance à être plus faibles que celles obtenues au moyen de l'estimateur ordinaire. Pour les chômeurs, en Ontario, l'écart entre les deux types d'estimation a tendance à être nettement plus faible.

Nous recherchons à l'heure actuelle un moyen de réduire les écarts systématiques entre les estimations pour des groupes de renouvellement différents. Plus précisément, nous étudions la possibilité de faire une correction de la pondération pour le nombre de ménages de tailles différentes selon le groupe de renouvellement pour tenir compte du fait que les ménages de petite taille sont

sous-représentées dans le nouvel échantillon. Une telle méthode donnerait de bons résultats pour les estimateurs composites et l'estimateur de régression ordinaire, et atténuerait probablement l'écart entre ces estimateurs.

3.1 Résultats empiriques

Emploi et chômage au niveau provincial. Le graphique 1 montre l'emploi total au niveau provincial de 1987 à 1998 pour l'Ontario. La série chronologique d'estimations composites pour les quatre valeurs de α , c'est-à-dire pour 0,6, 0,67, 0,75 et 1, se comporte de façon similaire. Ces graphiques montrent clairement qu'il y a une variation du niveau pour cette série dans le cas de l'estimation composite — le nombre estimé de personnes occupées est plus faible. Les versions désaisonnalisées de la série de données sur l'emploi en Ontario fondées sur l'estimateur ordinaire GREG et sur l'estimateur composite pour $\alpha = 0,67$ sont illustrées au graphique 2.

Le graphique 3 donne une comparaison de l'estimation ordinaire GREG du chômage en Ontario à l'estimation composite par régression pour $\alpha = 0,67$. L'effet de l'estimation composite sur cette variable est manifestement moins prononcée que sur les variables liées à l'emploi. Le graphique 4 donne une comparaison de la variation annuelle de l'emploi en Ontario pour les deux estimateurs. Chaque point de la série correspond à l'écart entre l'emploi durant le mois *m* de l'année *y* et le mois *m* de l'année *y* - 1. Par exemple, le premier point est le chiffre d'emploi en janvier 1988 dont est soustrait l'emploi en janvier 1987. La série d'estimations composites est de toute évidence plus lisse, particulièrement pour la deuxième moitié de la période de 12 mois.

Emploi selon la région intraprovinciale. Le graphique 5 donne une comparaison de l'estimation ordinaire et de l'estimation composite de l'emploi pour $\alpha = 0,67$ pour une région économique de l'Ontario. Les résultats pour d'autres régions intraprovinciales sont comparables. Le comportement des séries d'estimations ordinaires et composites étant fort semblable, l'effet de l'estimation composite n'est ni bénéfique ni néfaste. Pour les totalisations spéciales, le système de calcul des estimations de l'EPA est suffisamment souple pour permettre à l'utilisateur d'ajouter des valeurs de contrôle au niveau de la région économique, au besoin. Il existe déjà un contrôle pour l'estimation de la population totale selon la région économique.

Emploi selon la branche d'activité et désaisonnalisation. Nous avons comparé les estimations composites aux estimations ordinaires par régression pour 16 branches d'activité. Les graphiques 6A à 6D montrent les résultats pour deux d'entre elles en Ontario. Bien qu'elles ne soient pas incluses dans ces graphiques, de nouveau, les quatre valeurs de α produisent des séries d'estimations composites qui se comportent en général de façon comparable, même si la série obtenue pour $\alpha = 1$ s'écarte

graphiques et numériques sont présentes plus loin, à la section 3.1.

Application dans les systèmes. Un avantage important de l'estimateur tient au fait qu'il peut être appliqué dans l'ancien système d'estimation de l'EPA de façon simple, puisque, comme nous l'avons décrit plus haut, il suffit essentiellement d'agrandir la matrice de régression. Ce facteur a joué un rôle important dans notre décision d'étudier et, en bout de ligne, d'adopter l'estimation composite, car autrement, la transformation du système aurait coûté beaucoup plus cher.

Pondération. Dans le cas de l'estimateur A-K, la pondération pour satisfaisante aux totaux de population selon le groupe âge-sexe et la région géographique de contrôle. À titre d'illustration, nous décririons à l'annexe 3 la façon dont la matrice de régression serait agrandie si on ajoutait les éléments $x_i^{(C)}$ définis à la section 2. L'ajout des éléments $x_i = (1 - \alpha)x_i^{(L)} + \alpha x_i^{(C)}$ est similaire. Cette façon de procéder assure non seulement la cohérence des estimations décrites au paragraphe suivant, mais retient aussi les avantages qu'offrent les contrôles appliqués à l'estimateur de régression habituel, c'est-à-dire, ici, les contrôles selon le groupe âge-sexe et la région géographique.

Cohérence. Les estimations demeurent cohérentes, puisque la pondération en fonction des totaux de contrôle selon le groupe âge-sexe et la région géographique a lieu en même temps que la pondération en fonction des valeurs de contrôle pour l'estimation composite. En particulier, la somme des parties est égale au total; par exemple, personnes occupées + chômeurs = population active. Dans le cas d'autres méthodes d'estimation composite, la cohérence est assurée par d'autres moyens qui nécessitent une étape distincte ou une certaine forme de compromis.

Gains d'efficacité. Pour les variables ajoutées comme totaux de contrôle, la méthode produit des gains d'efficacité considérables aussi bien dans le cas de l'estimation du niveau que de la variation. Si $\alpha = 1$, pour les estimations de la variation, les gains d'efficacité peuvent être spectaculaires; en choisissant une valeur plus faible de α , nous obtenons un gain plus important pour les estimations du niveau, mais nous réduisons l'importance du gain pour les estimations de la variation. Certains résultats obtenus en choisissant $\alpha = 2/3$ sont présentés à la section 3.1.

Désaisonnalisation. Les séries chronologiques de données sur l'emploi pour diverses branches d'activité sont examinées de près par les utilisateurs tant internes qu'externes des données provenant de l'Enquête sur la population active. L'un des avantages importants du gain d'efficacité susmentionné est que plusieurs séries que l'on ne pouvait désaisonnaliser auparavant peuvent l'être

dans le programme d'assurance-emploi de l'administration fédérale). Fuller et Rao (2001) ont réservé ces contrôles L et C dans la régression, mais le grand nombre de colonnes que nous obtiendrions dans la matrice X aurait notamment pour conséquence indésirable d'augmenter le nombre de poids finals extrêmes, y compris des poids négatifs. Pour éviter ceci, nous serions obligés de limiter le nombre de branches d'activité incluses dans l'estimateur. Wayne Fuller (voir Fuller et Rao 2001) a proposé une autre solution qui permet d'inclure les branches d'activité présentant le plus d'intérêt, tout en faisant un compromis entre l'amélioration des estimations du niveau, d'une part, et de la variation, d'autre part. Sa solution consiste à prendre une combinaison linéaire de la colonne L et de la colonne C pour une branche d'activité donnée et de s'en servir comme nouvelle colonne dans la matrice X, c'est-à-dire à utiliser

$$x_i = (1 - \alpha)x_i^{(L)} + \alpha x_i^{(C)}.$$

Les variables originales axées sur le niveau ou sur la variation représentent des cas spéciaux du compromis de Fuller.

Choix de α : Fuller et Rao (2001) ont montré, en s'appuyant sur certaines hypothèses raisonnables que des valeurs de α telles que 0,65 et 0,75 produisent des estimations satisfaisantes tant du niveau que de la variation. Le choix de la valeur de α dépend de la variable étudiée (plus précisément de sa corrélation dans le temps) et de l'importance relative du niveau par rapport à la variation. Nos études (voir l'annexe 1) montrent que, pour les deux variables les plus importantes, c'est-à-dire les personnes occupées et les chômeurs, les meilleures valeurs de α pour les estimations du niveau sont 0,39 et 0,24, respectivement. Les valeurs correspondantes pour l'estimation de la variation sont 0,99 et 0,81, respectivement. Il faut donc maintenir un compromis entre l'amélioration des estimations du niveau et des estimations de la variation.

Pour déterminer quelles valeurs de α en calculant pour chaque variable la moyenne des valeurs basées sur le niveau et des valeurs basées sur la variation, nous avons obtenu ainsi les valeurs approximatives de 0,7 et 0,52 pour les personnes occupées et les chômeurs, respectivement. Les résultats fondés sur les valeurs de $\alpha = 1$ et $\alpha = 0,75$ ayant déjà été obtenus, nous y avons ajouté les résultats pour $\alpha = 0,67$ et $\alpha = 0,6$. Compte tenu des résultats présentés plus loin, qui n'indiquent aucune différence appréciable que l'on utilise 0,6, 0,67 ou 0,75 comme valeur de α , nous avons décidé d'utiliser la valeur $\alpha = 2/3$ dans le système de production.

3. CARACTÉRISTIQUES, PROPRIÉTÉS ET

RÉSULTATS

Nous résumons ici certaines caractéristiques et propriétés de l'estimateur composite de régression. Certains résultats

Si l'on s'intéresse principalement à l'estimation de la variation X qui suit donne de bons résultats :

$$x_{i,c}^{(c)} = \begin{cases} y_{i,c} & \text{si } i \in U \\ y_{i,c} + R(y_{i,c-1} - y_{i,c}) & \text{si } i \in M, \end{cases}$$

où R est un rapport de correction qui permet de tenir compte du fait que les cinq sixièmes de l'échantillon sont communs de mois en mois. La valeur $R = \sum w_i^{\text{tot}} / \sum w_i^{\text{apparié}}$ est celle qui est utilisée dans le système de production. Par souci de commodité, nous utilisons ici $R = 6/5$, puisque, en pratique, la différence entre les deux est faible, car on applique des méthodes visant à équilibrer les poids par groupe de renouvellement (par exemple, la correction pour la non-réponse est faite séparément, selon le groupe de renouvellement). Comme précédemment, le total de contrôle correspondant est l'estimation du nombre de personnes employées en agriculture le mois précédent. Par application des poids finals aux éléments de la colonne de la matrice X et totalisation, nous obtenons l'égalité

$$Y_{i-1}^{\prime} = Y_i^{\prime} - \Delta_{M,f}^{\prime}$$

ce qui signifie, en toutes lettres, que l'estimation du mois précédent est égale à la différence entre l'estimation du mois courant et une estimation Δ de $X_i - X_{i-1}$ fondée sur l'application des poids finals après l'estimation composite. Pour ce qui est des poids «avant l'estimation composite», il est facile de démontrer que, dans le cas d'une seule variable :

$$Y_i^{\prime} = (1 - b) Y_i + b (Y_{i-1}^{\prime} + \Delta_{M-1,i}^{\prime}),$$

où b est le coefficient de régression et Δ est l'estimation de la variation fondée sur les poids de départ. Le cas plus général où des variables auxiliaires sont incluses est décrit par Fuller et Rao (2001, équation 2.3).

Des études antérieures montrent que l'utilisation de contrôles axés sur le niveau (*level*), appelés ici contrôles L , produit de bonnes estimations de la variation pour les variables ajoutées à la matrice. Singh, et coll. (1997, 2001) décrivent les gains d'efficacité observés pour les estimations du niveau et de la variation axées sur la variation (contrôles C), et renvoient à des résultats antérieurs sur les estimations axées sur le niveau (contrôles L).

Au départ, il avait été proposé d'élaborer un système d'estimation qui s'appuierait uniquement sur des contrôles axés sur la variation (C). Cependant, certains auteurs ont exprimé des réserves à l'idée d'utiliser un tel système, puisqu'il est importantes (elles jouent, par exemple, un rôle essentiel

utiliser la régression pour la pondération. Chaque personne

qui figure dans l'échantillon correspond à une ligne de la matrice X . Chaque colonne de X correspond à un total de contrôle; par exemple si la colonne c correspond aux hommes de 20 à 24 ans, la valeur qui figure à l'intersection de la ligne i et de la colonne c sera égale à 1 si la personne i est un homme de 20 à 24 ans et sera égale à 0 autrement (la situation est similaire pour les colonnes qui correspondent aux régions géographiques). Pour des renseignements plus détaillés sur les méthodes d'estimation utilisées pour l'Enquête sur la population active, consulter Gambino, Singh, Dufour, Kennedy et Lindeyer (1998).

Pour tirer parti de l'échantillon qui est commun d'un mois à l'autre, on ajoute à la matrice X des colonnes dont les éléments sont définis de façon que, lorsque l'on applique les poids finals du mois courant aux éléments de chaque nouvelle colonne, le total soit une estimation composite basée sur le mois précédent, c'est-à-dire que l'estimation composite du mois précédent soit utilisée comme total de contrôle (strictement parlant, le total de contrôle est fondé sur les poids qui reflètent la population du mois courant). Comme nous l'avons mentionné dans l'introduction, il existe plusieurs moyens de définir les nouvelles colonnes, selon les objectifs poursuivis. Nous ne présentons plus bas que les solutions dont la mise en œuvre a été proposée.

Une nouvelle colonne type correspond à l'emploi dans une branche d'activité donnée, disons l'agriculture. Si l'on s'intéresse principalement aux estimations de nouveau, la méthode qui suit proposée pour former les colonnes donne de bons résultats. Représentons par M et U l'échantillon apparié (commun) et non apparié (nouvelles unités d'échantillonage), respectivement. Pour la personne i , aux périodes $t-1$ et t , représentons par $y_{i,t-1}^{\prime}$ et $y_{i,t}^{\prime}$ les variables indicateurs dont la valeur est égale à 1 chaque fois que la personne travaille en agriculture. Alors, posons

$$x_{i,c}^{(c)} = \begin{cases} y_{i,t-1}^{\prime} & \text{si } i \in U \\ y_{i,t-1}^{\prime} & \text{si } i \in M, \end{cases}$$

où $y_{i,t-1}^{\prime}$ représente l'estimation composite de la proportion de personnes qui travaillaient en agriculture le mois précédent; en pratique, nous utilisons $y_{i,t-1}^{\prime} = Y_{i,t-1}^{\prime} / P_{15+}$, où P_{15+} représente la population de 15 ans et plus. Le total de contrôle correspondant est l'estimation du nombre de personnes qui travaillaient en agriculture le mois précédent, c'est-à-dire Y_{t-1}^{\prime} . Donc, en dernière analyse, la somme pondérée finale des éléments de la nouvelle colonne sera égale à l'estimation du mois précédent. Cette situation équivaut presque à forcer les poids du mois courant, appliqués aux valeurs du mois précédent pour l'échantillon commun, à reproduire l'estimation de l'emploi en agriculture du mois précédent (après l'application d'un facteur de correction de 5/6). Nous utilisons l'exposant L pour rappeler que l'objectif est ici d'améliorer les estimations de niveau (*level*).

1990, Statistique Canada a recommencé à s'intéresser à l'estimation composite et à mis au point une méthode par régression qui s'adapte bien au système existant de calcul des estimations de l'EPA. Cette méthode est décrite dans Singh, Kennedy, Wu et Brisebois (1997), et une version plus à jour est présentée dans Singh, Kennedy et Wu (2001). La nouvelle méthodologie permet de faire un choix entre diverses méthodes, selon l'objectif poursuivi. Si l'on s'intéresse principalement à l'estimation du niveau, on peut utiliser les prédicteurs axés sur le niveau. Par contre, si l'on accorde plus d'importance à la variation, on peut utiliser les prédicteurs axés sur la variation. On peut aussi aller une étape plus loin et inclure les deux types de prédicteurs dans la méthode. Cependant, dans ce dernier cas, le grand nombre de variables indépendantes dans l'équation de régression peut produire une distorsion des poids d'échantillonnage finaux.

Les résultats provisoires obtenus d'après la nouvelle méthode en se servant des prédicteurs axés sur la variation et d'autres obtenus en se servant des prédicteurs axés sur le niveau ont été examinés à la réunion du Comité consultatif des méthodes statistiques de Statistique Canada. La méthode permet de résoudre les problèmes que posent les estimateurs composites et de réaliser des gains d'efficacité considérables. Toutefois, l'estimateur qui utilise les prédicteurs axés sur la variation peut causer une dérive des estimations du niveau au fil du temps dans certaines situations extrêmes. Enfin, il a été décidé, à la suite des recommandations du Comité, de donner, tant pour les estimations du niveau que de la variation, un ordre d'importance pour le choix des prédicteurs. Après l'échange de notes techniques entre Wayne Fuller, J.N.K. Rao et les employés de Statistique Canada, le Bureau a adopté une méthode proposée par Fuller, qui combine les stratégies axées sur le niveau et sur la variation sans poser les contraintes dues à l'intégration des deux ensembles de prédicteurs dans la régression (voir Fuller et Rao 2001). La solution est étonnamment simple : prendre une combinaison linéaire des prédicteurs de niveau et de variation : $X = (1 - \alpha)X_T + \alpha X_C$, et l'utiliser comme prédicteur. Les prédicteurs de variation et de niveau deviennent alors des cas spéciaux. De surcroît, on peut choisir α de façon à refléter l'importance relative que l'on veut accorder à l'estimation du niveau par opposition à l'estimation de la variation.

Le présent article décrit le nouvel estimateur composite à la section 2. On a procédé à l'évaluation approfondie de cet estimateur en se servant de données réelles de l'EPA pour plusieurs caractéristiques observées pendant une longue période. Les résultats de ces études sont résumés à la section 3. Contrairement aux estimateurs composites classiques, l'estimateur composite de régression exige que l'appariement de l'échantillon entre deux mois consécutifs se fasse au niveau de l'enregistrement individuel. Cette contrainte produit des situations intéressantes où il faut traiter les non-répondants, ainsi que les personnes dans le

champ et hors du champ d'un mois à l'autre de façon à ce que la qualité des estimations de la variation ne soit pas compromise. À la section 4, on examine la méthode d'imputation mise au point pour résoudre les divers problèmes que posent des données qui manquent pendant deux mois consécutifs. Enfin, l'évaluation de la qualité de ce nouvel estimateur composite se fonde non seulement sur son efficacité statistique, mais aussi sur sa stabilité au cours du temps, sa rentabilité et la capacité d'atteindre les objectifs suivants : i) réduire au minimum les changements à apporter à l'ancien système d'estimation, ii) produire un poids unique pour chaque unité d'échantillonnage, iii) produire des estimations qui concordent avec les totaux de contrôle selon la catégorie âge-sexe et la région géographique et iv) produire des estimations cohérentes (au sens où, par exemple, personnes occupées + chômeurs = population active et population active + population inactives = personnes de 15 ans et plus). Ces objectifs sont examinés à divers points dans l'article, mais plus spécialement à la section 3.

2. L'ESTIMATEUR COMPOSITE DE RÉGRESSION

Certaines enquêtes, notamment la *Current Population Survey* (CPS) aux États-Unis, mettent à profit le chevauchement de l'échantillon grâce à l'utilisation des estimateurs composites K ou AK . À l'origine, dans le cas de la CPS, on utilisait l'estimateur composite K

$$y'_t = (1 - K)y_t + K(y'_{t-1} + \text{variation}_{t-1, \mu})$$

où $K = 1/2$ à la période t , et où variation $_{t-1, \mu}$ représente une estimation de la variation fondée sur l'échantillon commun, ou apparié. Cet estimateur a été remplacé par la suite par l'estimateur composite AK

$$y'_t = (1 - K)y_t + K(y'_{t-1} + \text{variation}_{t-1, \mu}) + A(\text{non apparié} - \text{apparié})$$

où $A = 0,2$ et $K = 0,4$ (voir Cantwell et Ernst 1992). Les valeurs optimales de A et K dépendent de la variable étudiée; or, si l'on applique ce dernier estimateur, l'utilisation de valeurs différentes pour des variables différentes pose des problèmes de cohérence (en ce sens que la somme de parties n'est pas égale au total). Ces problèmes nous ont poussés à rechercher d'autres méthodes pour répondre aux objectifs mentionnés à la fin de la section précédente.

Il convient de souligner que nous décrivons ici l'application de la nouvelle méthode au niveau des particuliers, mais qu'en pratique, les données au niveau des particuliers sont agrégées au niveau du ménage et que ce sont donc les enregistrements concernant les ménages qui sont utilisés pour produire les estimations.

Dans le cas de l'ancien système de calcul des estimations de l'EPA, on crée une matrice de régression X pour pouvoir

Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application

JACK GAMBINO, BRIAN KENNEDY et MANGALA P. SINGH¹

RÉSUMÉ

L'Enquête sur la population active (EPA) du Canada est une enquête mensuelle réalisée selon un plan de sondage complexe avec renouvellement de panel. Après des études approfondies, y compris l'examen de diverses méthodes tirant parti du chevauchement de l'échantillon pour améliorer la qualité des estimations, l'équipe de l'EPA a choisi une méthode d'estimation composite qui permet d'atteindre cet objectif tout en respectant les contraintes d'ordre pratique. En outre, pour les variables pour lesquelles le gain d'efficacité est important, la nouvelle série chronologique a tendance à être plus compréhensible pour les spécialistes du domaine. Il est donc plus facile d'expliquer les estimations produites d'après l'EPA aux utilisateurs des données et aux membres des médias. Comme l'estimation composite réduit la variance, il est maintenant possible de publier des estimations mensuelles dans bien des cas où, jusqu'à présent, seules les moyennes mobiles de trois ans pouvaient être publiées. En outre, la méthode permet de désaisonnaliser correctement un grand nombre de séries chronologiques.

MOTS CLÉS : Enquête avec renouvellement de panel; système d'estimation; pondération; estimation de la variance; estimation du niveau.

1. INTRODUCTION

1.1 Pourquoi recourir à l'estimation composite?

L'Enquête sur la population active (EPA) du Canada est une enquête mensuelle réalisée après de 54 000 ménages sélectionnés selon un plan d'échantillonnage stratifié à plusieurs degrés. Comme les nouveaux ménages sélectionnés sont ajoutés à l'échantillon pendant six mois consécutifs, les cinq sixièmes de l'échantillon sont communs d'un mois à l'autre. Chaque mois, on pose aux membres des ménages faisant partie de l'échantillon des questions sur la situation d'activité, les gains et d'autres aspects de l'activité. Dans le système d'estimation de l'EPA utilisé avant 2000, les poids de sondage initiaux étaient modifiés par régression pour produire des poids finals obéissant aux *taux de population* selon la catégorie âge-sexe et la région géographique (niveau infraprovincial) utilisés comme contrôles. On obtenait ainsi pour chaque enregistré un *poids final unique* que l'on utilisait pour toutes les totalisations.

Ce système d'estimation n'utilisait que les données du mois courant et ne tirait nullement parti du fait que l'on peut se servir de l'échantillon commun pour améliorer les estimations. Pourtant, certaines caractéristiques, comme l'emploi selon la branche d'activité, sont fortement corrélées dans le temps et d'autres, comme le chômage, le sont modérément, si bien que des gains d'efficacité seraient possibles. C'est pour cette raison que, pour des enquêtes comparables à l'EPA, comme la *Current Population Survey* aux États-Unis, on recourt depuis des années à l'estimation composite pour améliorer les estimations. Cependant, le

Au début des années 1980 (voir Kumar et Lee 1983), la stratégie d'estimation composite de la CPS a été étudiée en vue de son application éventuelle à l'EPA. Bien que cette étude ait montré que l'estimation composite augmentait l'efficacité de l'estimation de l'emploi et, dans une moindre mesure, de l'estimation du chômage, on a estimé à l'époque que les aspects négatifs de la méthode surpassaient les gains. Entre autres, le fait que les paramètres qui optimisent l'estimation de l'emploi et du chômage soient assez différents aurait obligé à trouver un équilibre entre, d'une part, l'utilisation d'un ensemble de paramètres inter-médiaires, qui réduirait les gains d'efficacité et, d'autre part, l'obtention de valeurs des variables dont la *somme n'est pas égale aux totaux* (par exemple, la somme des personnes occupées et des chômeurs ne serait pas égale à la population active, à moins que l'une des trois valeurs ne soit obtenue par différence). Le fait que cette méthode d'estimation composite n'était pas compatible avec les *systèmes de pondération, d'estimation et de diffusion* utilisés à ce moment-là pour l'EPA était un autre facteur en sa défaveur, puisque son adoption aurait nécessité un remaniement complet des systèmes.

Après avoir, les taux de chômage mensuels étaient les principales estimations produites d'après l'Enquête sur la population active. Toutefois, comme l'intérêt pour l'estimation du taux d'emploi et de sa variation s'est accru ces dernières années, il a fallu multiplier les efforts en vue de découvrir un moyen de profiter de l'échantillon commun, puisque la production de ces estimations en bénéficierait considérablement. Par conséquent, au milieu des années

¹ Jack Gambino, Brian Kennedy et Mangala P. Singh, Statistique Canada, Parc Tunney, Ottawa, (Ontario), Canada K1A 0T6.

BIBLIOGRAPHIE

- LENT, J., MILLER, S. et CANTWELL, P. (1994). Composite weights for the Current Population Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 867-872.
- LENT, J., MILLER, S. et CANTWELL, P. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 130-139.
- RAO, C.R. (1973). *Linear Statistical Inference and its Applications*. Deuxième édition. New York : John Wiley and Sons.
- SINGH, A.C., and MERKOURIS, P. (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.
- SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 120-129.
- SINGH, A.C., KENNEDY B., WU S. et BRISEBOIS F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.
- SHISKIN, J., YOUNG, A., and MUSGRAVE, J. (1967). *The X-11 variant of Census Method II Seasonal Adjustment*, Bureau of the Census, U.S. Department of Commerce, Paper technique 15.
- YANSANEH, I.S., et FULLER, W.A. 1998. Méthode optimale d'estimation réursive pour les enquêtes répétitives. *Techniques d'enquête*, 24, 33-42.
- KOTT, P.S. (1998). Using the delete-a-group jackknife variance estimator in practice. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 483-486.
- BAILLAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of American Statistical Association*, 70, 23-29.
- BELL, P.A. (1998). Using state space models and composite estimation to measure the effects of telephone interviewing on labour force estimates. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 98/2, ABS, Canbera.
- BELL, P.A. (1999). The impact of sample rotation patterns and composite estimation on survey outcomes. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 99/1, ABS, Canbera.
- BELL, P.A., et CAROLAN, A. (1998). Trend estimation for small areas from a continuing survey with controlled sample overlap. Document de travail *Econometrics and Applied Statistics*, numéro de 1351.0, no. 98/1, ABS, Canbera.
- FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.
- FULLER, W.A. (1999). Canadian Regression Composite Estimation. Manuscript non publié.
- GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.
- JESSEN, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts. *Iowa Agricultural station research Bulletin*, 304.

Le tableau 1 présente ces erreurs-types moyennes pour diverses mesures de tendance et désaisonnalisées, relativement à celles que l'on peut obtenir de l'estimateur RG courant, tant pour les personnes ayant un emploi que pour les personnes sans emploi. Dans ce même tableau on trouve les chiffres correspondants pour le niveau, le mouvement, la moyenne trimestrielle et le mouvement de la moyenne trimestrielle, en fonction des diagrammes 5 à 8.

l'estimateur RG courant

Personnes ayant un emploi :	Niveau	93	92	89	82	83
	Mouvement	95	95	89	66	69
	Moyenne trimestrielle	93	92	89	85	85

Niveau
Mouvement
Moyenne trimestrielle

Personnes sans emploi :	
Niveau	100
Mouvement	101
Moyenne trimestrielle	100
Mouvement de la moyenne trimestrielle	97
Désaisonnalisée	100
Mouvement de la désaisonnalisation	102
Tendance de fin	99
Mouvement de la tendance de fin	97
Révision de la tendance	95
Révision du mouvement de la tendance	97

J'estime que, pour de nombreuses applications, les indicateurs les plus importants sont ceux qui indiquent la

REMERCIEMENTS

REMERCIEMENTS

6.6 Résumé

direction sous-jacente de la série à la fin courante, c'est-à-dire le mouvement de la moyenne trimestrielle et le mouvement de la tendance. Une réduction de l'erreur-type pour ces éléments rend la direction sous-jacente de la série à la fin plus claire, même pour les utilisateurs qui se fient sur une inspection visuelle ou sur un processus de lissage quelconque autre que la tendance X11. Cela offre dès lors de meilleures chances de détecter les points tournants de la série sous-jacente.

Pour ce qui est du mouvement de la tendance, l'estimateur B1 permet une réduction de 18 % de l'erreur-type pour les personnes ayant un emploi et une réduction de 8 % pour les personnes sans emploi. Quant au MR2, ces réductions sont de 35 % et de 7 % respectivement. Les estimateurs composites permettent également de réduire l'apport de l'erreur d'échantillonnage aux révisions de la série des tendances.

ET : % de l'ER courante

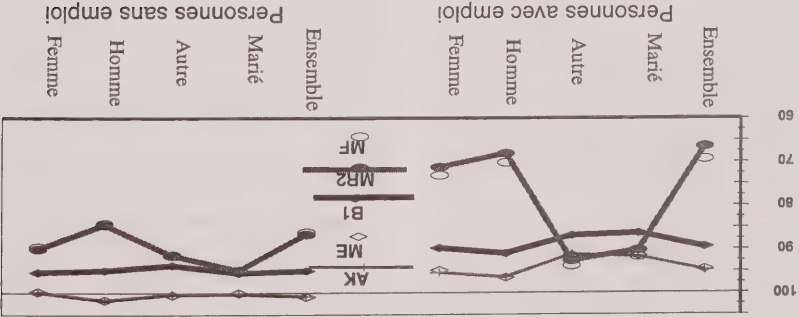


Diagramme 6. Erreur-type du mouvement (% de l'ET courante)

ET : % de l'ER courante

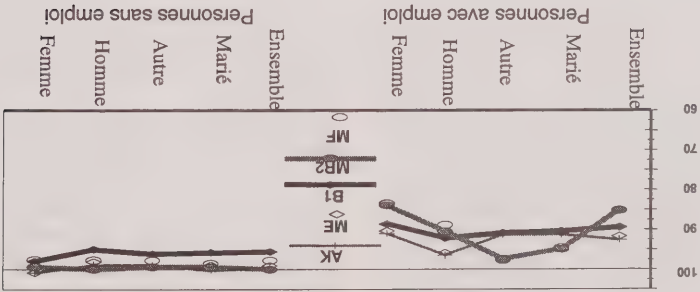


Diagramme 7. Erreur-type de la moyenne trimestrielle (% de l'ET courante)

ET : % de l'ER courante

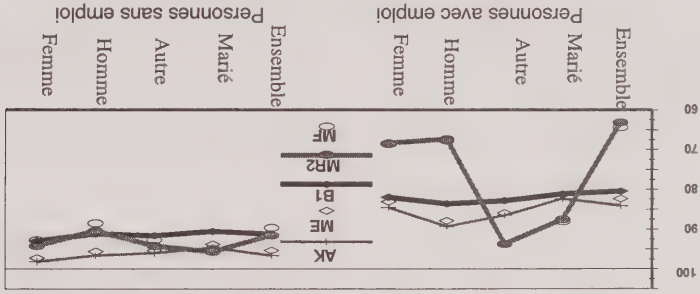


Diagramme 8. Erreur-type du mouvement de la moyenne trimestrielle (% de l'ET courante)

La valeur de la tendance pour tel moment est révisée à mesure que les données pour des moments ultérieurs sont accessibles. J'ai estimé l'erreur-type des estimations de tendance à la fin de la série (tendance de fin) et pour les mêmes moments lorsque douze autres mois de données sont disponibles (tendance à mi-chemin). Les révisions de la tendance (ou du mouvement de tendance) sont définies comme la différence entre la valeur à mi-chemin et la valeur de fin de la tendance (ou du mouvement de tendance). L'ampleur de la révision dépend de la forme de la série réelle et de l'erreur d'échantillonnage des séries estimées. La révision de la tendance quadratique moyenne pour une série d'estimations non biaisées est la somme de deux composantes : la révision de la tendance quadratique

La valeur de la tendance pour tel moment est révisée à mesure que les données pour des moments ultérieurs sont accessibles. J'ai estimé l'erreur-type des estimations de tendance à la fin de la série (tendance de fin) et pour les mêmes moments lorsque douze autres mois de données sont disponibles (tendance à mi-chemin). Les révisions de la tendance (ou du mouvement de tendance) sont définies comme la différence entre la valeur à mi-chemin et la valeur de fin de la tendance (ou du mouvement de tendance). L'ampleur de la révision dépend de la forme de la série réelle et de l'erreur d'échantillonnage des séries estimées. La révision de la tendance quadratique moyenne pour une série d'estimations non biaisées est la somme de deux composantes : la révision de la tendance quadratique

La technique jackknife avec suppression d'un groupe a permis de préparer des estimations de l'erreur-type pour les différentes estimations de tendance et désaisonnalisées. Cette technique suppose la préparation de versions répétées

septembre 1997.

ment apparues ne manifeste pas cet important biais

manifestent un certain caractère saisonnier, mais elles sont

diverses estimations par rapport à l'estimation RG pour les

marqué que pour le passage aux autres estimateurs.

6.4 Erreurs-types

pourcentages pour la période en fonction de laquelle ils ont

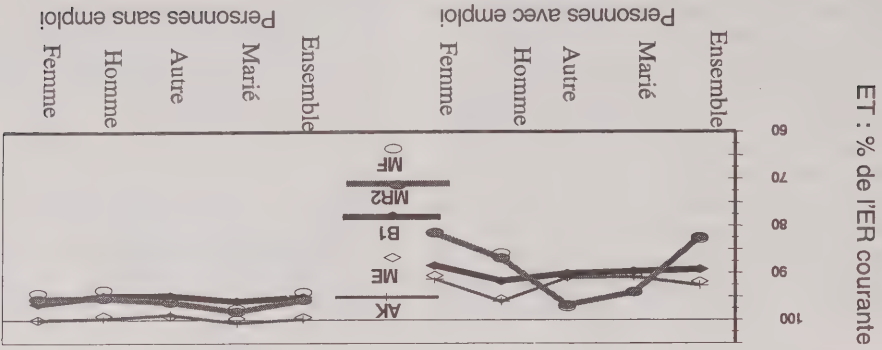


Diagramme 5. Erreur-type du niveau (% de l'ET courante)

pour les personnes ayant un emploi.

6.5 Série désaisonnalisée et série de tendances

MF que pour l'estimateur B1.

la présente étude ne semblait pas pouvoir concurrencer

ment, de la moyenne trimestrielle et du mouvement de la

Les estimations AK, ME et B1 sont assez semblables, puisque, pour l'ensemble des trois méthodes, l'apport d'une unité dépend de son groupe de renouvellement. Dans les deux diagrammes, les estimateurs AK, ME et B1 semblent donner, en moyenne, des valeurs plus faibles que les estimations RG. Cela indique un changement du biais d'accoutumance, résultat du fait que l'on accorde un poids moindre au groupe de renouvellement échantillonné pour la première fois. Les estimations varient vers le haut et vers le bas relativement à leur différence moyenne pour de courtes périodes.

Les estimations MR2 et MF tendent à différer des autres estimations, puisqu'elles soulignent l'apport des unités de l'échantillon séparé. Pour les personnes ayant un emploi, les estimateurs MR2 et MF sont appréciablement plus grands en moyenne que les estimations RG, jusqu'à septembre 1997. Il y a alors une baisse des différences correspondant à l'intégration d'un nouvel échantillon à partir de septembre 1997. Pour des raisons qui ne sont pas claires, au cours de cette période l'échantillon apparaît manifeste un comportement différent de celui de l'échantillon global. Cela influe sur la différence entre ces séries

Afin de quantifier le changement probable de biais suivant le passage à un nouvel estimateur, on a calculé la différence moyenne au cours de la période de chaque estimation relativement à l'estimation RG. Il est possible que la différence soit saisonnière, et l'on a donc obtenu des moyennes distinctes pour chaque mois de l'année civile, de même que pour l'ensemble. On trouvera dans le diagramme 3 les différences moyennes au cours de la période allant de juillet 1993 à janvier 1999 pour les personnes ayant un emploi.

6.3 Différences moyennes selon le mois civil

Pour les personnes sans emploi, les estimations MR2 et MF tendent à être inférieures aux estimations RG. Il n'y a aucune indication d'un « problème de dérive » pour les personnes sans emploi, ce qui n'a rien de surprenant compte tenu des corrélations moindres mises en jeu.

de dérive ».

Le niveau de la série MR2 au cours d'une période subséquente considérable, manifestation possible du « problème

Diagramme 3. Différence moyenne par rapport à l'estimation RG, globalement et selon le mois civil, personnes ayant un emploi (milliers).

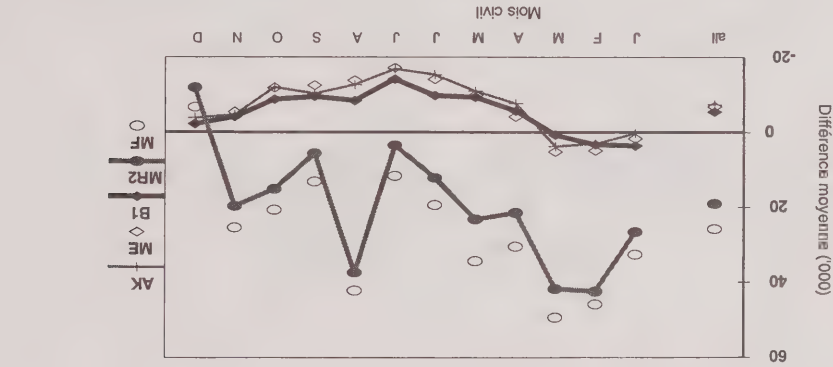
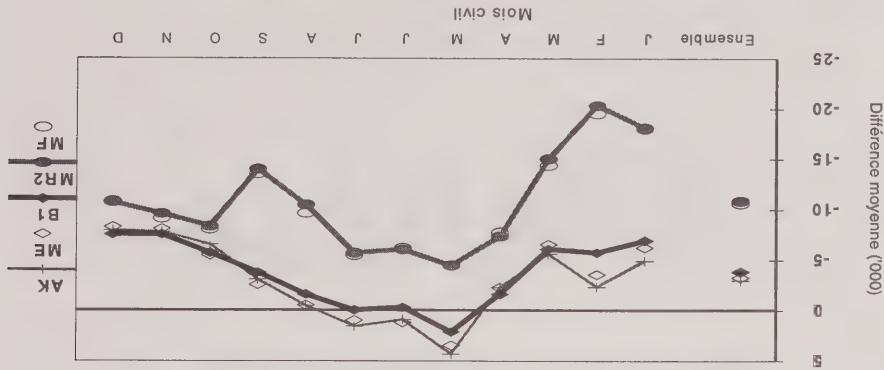


Diagramme 4. Différence moyenne par rapport à l'estimation RG, globalement et selon le mois civil, personnes sans emploi (milliers).



préparer les estimations de régression modifiée du présent rapport, des variables z ont été préparées pour les estimations de l'emploi et du chômage pour chaque État et chaque sexe. Cela donne au total 32 variables auxiliaires supplémentaires, en plus des 540 repères de strate après sélection habituels utilisés dans la régression généralisée.

6.2 Différences par rapport à l'estimation RG

On peut utiliser l'estimateur RG courant comme base de comparaison pour les autres estimateurs. Au lieu de présenter des diagrammes d'estimations de niveau, je montre les différences des autres estimations par rapport aux estimations RG courantes. Les diagrammes 1 et 2 indiquent ces différences pour les estimations des personnes ayant un emploi et des personnes sans emploi respectivement. Afin de situer l'ampleur de ces différences, notons que les erreurs-types publiées pour l'estimation courante étaient de 25 200 pour les personnes ayant un emploi et de 7 900 pour les personnes sans emploi en janvier 1999 (semblables pour les autres mois).

On a obtenu les estimations et les erreurs-types pour chacun des estimateurs ci-dessous (énumérés en abrégé pour faciliter la suite de l'exposé) :

RG : L'estimation de régression généralisée utilisée couramment dans l'EPA

AK : L'estimation AK avec $K=0,7$, $A=0,06$

ME : L'estimateur MELSB fondé sur une fenêtre de 7 mois

B1 : Le MELSB amélioré fondé sur une fenêtre de 7 mois

MR2 : L'estimateur MR2 (régression modifiée avec $a = 1$)

MF : La variante de Fuller de la régression modifiée

($a = 0,7$)

Les estimateurs de régression modifiée supposent un choix des variables clés qu'il s'agit d'optimiser. Pour

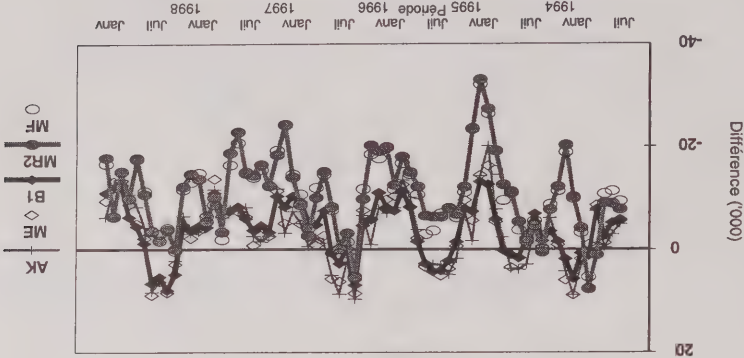
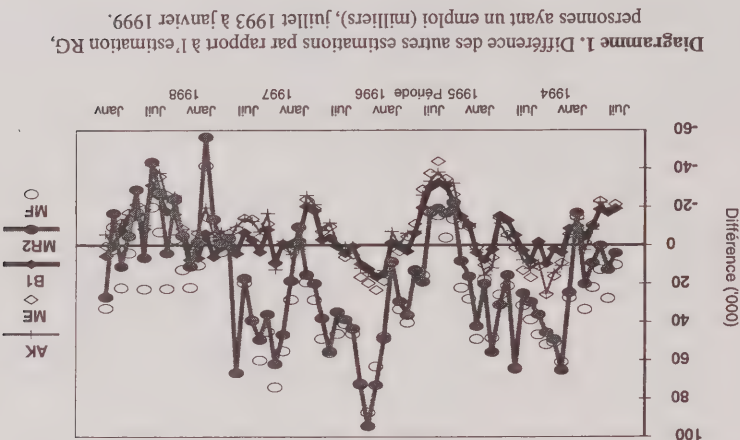


tableau est le même que pour la RG multipliée par la taille de la fenêtre. Il existe également une possibilité d'estimations négatives pour de petites cellules ne comportant aucune unité contrainte.

À noter que, dans l'estimateur B1, on ne force pas l'addition des poids appliqués aux mois autres que le mois courant de donner une valeur de zéro. En vertu des hypothèses du modèle, l'estimation y_{BH}^{st} et x_{BH}^{st} demeurent inconditionnellement sans biais, puisque y_{BH}^{st} et x_{BH}^{st} sont sans biais pour Y_t et X_t respectivement. Dans la pratique, le mois courant contribue 99,5 % environ du poids total. L'estime que le biais qui en résulte est faible et sans danger (entraînant un léger lissage des estimations avec le temps). Pour toute estimation dans laquelle il existe une corrélation appréciable des données d'un mois à l'autre, les estimations MELSB et B1 devraient comporter une erreur d'échantillonnage moindre que l'estimation RG. C'est là un avantage théorique relativement à une méthode servant à améliorer un ensemble d'estimations déterminée à l'avance (comme la régression modifiée ou l'AK avec pondération composite). Dans la pratique, cet avantage n'est pas nécessairement très important, car pour l'EPA on s'intéresse surtout à un petit nombre d'estimations bien définies.

L'utilisateur doit également déterminer la période de temps ou la « fenêtre » des estimations. Le fait d'utiliser trop de moments suppose beaucoup de temps d'ordinateur, tandis qu'un nombre trop faible réduit les avantages. La fenêtre de sept mois utilisée ici a été suffisante pour offrir presque tous les avantages ; des fenêtres plus petites comportent des erreurs-types appréciablement plus élevées.

6. COMPARAISON DES MÉTHODES

6.1 Méthode de comparaison

Les estimations pour la période allant de juillet 1993 à janvier 1999 ont été préparées en fonction de données allant de janvier 1993 à janvier 1999. On a obtenu des estimations classées selon le mois, l'État, le sexe, l'état matrimonial et la situation par rapport au marché du travail. On a également obtenu des estimations pour le mouvement de décalage un, la moyenne trimestrielle et le mouvement de moyennes trimestrielles successives.

L'erreur-type de ces estimations a été calculée à l'aide de la technique « jackknife avec suppression d'un groupe » (Kott 1998). Les unités géographiques qui constituent la première étape du tirage de l'échantillon ont été divisées systématiquement en $G = 30$ groupes, et « des groupes répétés » ont été établis en fonction de l'échantillon complet à l'exclusion des unités de l'un de ces groupes. Chaque estimation étudiée a également été préparée pour chacun des groupes répétés G . Lorsqu'on écrit e pour l'estimation de type jackknife avec suppression d'un groupe de l'erreur-type est donnée par

emploi, mais fonctionne bien aussi pour des personnes sans emploi. L'utilisation d'autres valeurs pour les paramètres du modèle n'a donné aucune amélioration appréciable de l'erreur-type pour les personnes sans emploi.

5.3 Estimations MELSB améliorées

Une difficulté liée aux estimations MELSB ci-dessus est qu'il faut des estimations RG au niveau du groupe de renouvellement. La taille plus petite de l'échantillon au niveau du groupe de renouvellement risque de limiter les répétitions utilisables, comme on l'a vu pour l'AK. Pour le MELSB, toutefois, une autre stratégie est possible.

On définit l'estimateur B1 en établissant un estimateur MELSB fondé sur les estimateurs de Horvitz-Thompson au niveau du groupe de renouvellement, puis en appliquant la technique de régression généralisée afin d'améliorer cet estimateur. Il s'agit de procéder comme suit. On définit $y_{BH}^{\#} = a_{IR(t,t)} x_{BH}^{\#}$ et $x_{BH}^{\#} = a_{IR(t,t)} x_{BH}^{\#}$, où $a_{IR(t,t)}$ est le multiplicateur MELSB applicable au groupe de renouvellement dans lequel se trouve l'unité i à un moment t . On peut alors écrire l'estimateur MELSB fondé sur les estimateurs de Horvitz-Thompson sous la forme

$$y_{BH}^t = \sum_{i=1}^{s-t-1} \sum_{\#} w_{\pi}^{st} y_{BH}^{st\#} \quad (19)$$

En calant en fonction des repères, on obtient l'estimateur MELSB amélioré B1 :

$$y_{BH}^t = y_{BH}^t + (X_t - x_{BH}^t) \beta \quad (20)$$

$$\text{pour } \hat{\beta} = \left(\sum_{i=1}^{s-t-1} \sum_{\#} w_{\pi}^{st} x_{BH}^{st\#} \right)^{-1} \sum_{i=1}^{s-t-1} \sum_{\#} w_{\pi}^{st} y_{BH}^{st\#} \quad (21)$$

$$\text{donc, } y_{BH}^t = \sum_{i=1}^{s-t-1} \sum_{\#} w_{\pi}^{st} y_{BH}^{st\#} \quad (22)$$

$$\text{pour } w_{BH}^{st} = w_{\pi}^{st} a_{SR(s,t)} \left\{ 1 + (X_t - x_{BH}^t) \right\} \quad (23)$$

Propriétés des estimateurs MELSB et B1

Les estimations de type MELSB et B1 sont des sommes de données unitaires pondérées pour une fenêtre de mois. Chaque estimation n'exige que des données de cette fenêtre et se laisse préparer indépendamment des estimations pour les mois précédents, de sorte que la méthode n'est pas réursive.

Le même mois de données contribuera différents poids à l'estimation pour différents moments. Une unité contribuera un poids appréciable à son estimation du mois courant, et un poids proche de zéro, souvent négatif, aux estimations d'autres mois. Le travail de préparation d'un

l'échantillon apparté. Ce peut être effectivement le cas : l'échantillon apparté exclut les personnes qui ont changé de logement entre les deux mois, et il est possible que le changement de logement soit lié à un changement d'emploi.

Ce « biais de l'échantillon apparté » vient s'ajouter à tout biais d'accoutumance.

Une autre difficulté est liée à l'estimateur MIR2 (donc, $a = 1$). Si la k -ième variable clé $y_{i,k}$ comporte une forte corrélation d'un mois à l'autre, elle le comportera également une forte corrélation avec la k -ième nouvelle variable auxiliaire $z_{i,k}$. Pour une telle variable, l'élément de $y_{i,k}^M$ qui correspond à $z_{i,k}$ aura une certaine valeur $y_{i,k}^M$ de un. Si l'on utilise (7), (11) et $Z_i \approx y_{i-1}^M$, l'estimateur MR2 adopte la forme

$$y_{i,k}^M \approx (1 - y_i) y_{i,k}^{*H} + y_i (y_{i-1,k}^{*H} - y_{i-1,k}^{*HD}) + (y_{i,k}^{*HD} - y_{i-1,k}^{*HD})$$

+ autres termes.

(14)

Dans ce cas, il est possible que le mouvement de l'échantillon apparté à un moment donne influence fortement les estimations lors de nombreux moments subséquents. De plus, un biais faible dans le mouvement aura tendance à s'accumuler avec le temps. Ce danger a été reconnu par Fuller (1999) et qualifié de « problème de dérive ». C'est ce qui l'a poussé à suggérer la forme d'estimateur donnée ici, avec une valeur de a inférieure à 1. En résumé, la régression modifiée offre des avantages semblables à ceux de la stratégie de pondération composite AK, mais avec une erreur d'échantillonnage possiblement inférieure. La méthode n'est pas difficile à appliquer, et elle permet d'éviter la nécessité d'un calage distinct des groupes de renouvellement en fonction des repères.

5. MEILLEURE ESTIMATION LINÉAIRE SANS BIAIS (MELSB)

5.1 MELSB pour une fenêtre fixe

On obtient l'estimateur MELSB pour une fenêtre fixe (noté y_B^i) en choisissant une combinaison linéaire « optimale » des estimations de groupe de renouvellement y_R^i (définie en 2.3) à partir d'une fenêtre de $l + 1$ mois, comme suit :

$$y_B^i = \sum_{l=1}^s a^{s-l} \sum_{r=1}^l y_R^i \quad (15)$$

où les paramètres a^{s^p} sont choisis de façon à réduire (y_B^i) au minimum moyennant les contraintes $\sum_{s=1}^l a^{s^p} = 1$ pour $s = 1, \dots, l$. Ces contraintes seront sans biais pour y_R^i , à la condition que les estimations de groupe de renouvellement soient sans biais, c'est-à-dire que $B(y_{Gr}^i) = Y^s$ pour $s = 1, \dots, l$. La réduction au minimum suppose que les variances et les covariances des estimations de groupe de renouvellement soient connues. Dans la pratique, celles-ci sont

estimées en fonction de données chronologiques. Le problème se laisse alors écrire sous une forme matricielle : nous cherchons à choisir le vecteur de colonne a (avec les éléments a^{s^p} pour $s = 1, \dots, l$ et $r = 1, \dots, 8$) de façon à réduire une forme quadratique $a'Va$ au minimum sous réserve des contraintes $C'a = c$. Le résultat standard pertinent (Rao 1973, page 65) est que le minimum survient pour $a = V^{-1}C^{-1}cq$ où q est une solution de $(C'V^{-1}C)q = c$. Dans la présente étude, la matrice V a été remplacée par une matrice de corrélation, suivant l'hypothèse selon laquelle toutes les estimations de groupe de renouvellement dans la fenêtre comportaient la même variance.

5.2 Structure de corrélation des estimations de groupe de renouvellement

Puisque différents modes de corrélation donnent diverses estimations MELSB, le choix d'un profil de corrélation comporte des aspects semblables à ceux du choix des paramètres A et K pour la stratégie composite AK. Il est souhaitable d'utiliser la même combinaison linéaire pour toutes les estimations afin de garantir l'additivité des estimations. J'ai adopté un modèle à quatre paramètres pour le mode de corrélation :

$$\text{corr}(y_{Gr}^i, y_{Gr}^s) = p_W^{l-s} \quad \text{pour } r - r' = l - s$$

$$= p_B^{l-s} \quad \text{pour } r - r' = l - s + 8m$$

pour la nombre entier $m \neq 0$

(16)

$= 0$ autrement.

Ainsi, la corrélation entre les estimations pour un décalage k du même groupe de renouvellement est p_k^w si le groupe de renouvellement comporte les mêmes logements aux deux moments, et p_k^B autrement. Il n'y a pas de corrélation entre les estimations de différents groupes de renouvellement. Un modèle à quatre paramètres est utilisé :

$$p_k^w = (1 - r_U^l)(\theta_k^p r_U^l + \theta_k^B(1 - r_U^l)) \quad \text{et} \quad (17)$$

$$p_k^B = (1 - r_U^l)\theta_k^B(1 - r_U^l) \quad (18)$$

Beil et Carolan (1998) ont discuté ce modèle. Les valeurs paramétriques utilisées dans cet exposé étaient $\theta_p = 0,87697$, $\theta_B = 0,94$, $r_U^l = 0,3101$ et $r_U^p = 0,90456$. Ces valeurs sont le résultat d'un ajustement du modèle en fonction d'autocorrélations estimatives pour les estimations de groupe de renouvellement de la proportion d'emploies. Il importe de noter que les estimations de type MELSB sont sans biais peu importe l'exactitude du modèle de corrélation supposé. Le modèle utilise ici cherche à être optimal pour des estimations des personnes ayant un

compris les estimations de personnes ne relevant pas de la population active) ne sont typiquement pas très améliorées relativement aux estimations RG standard (Lent, Miller et Cantwell 1996).

4. ESTIMATION DE RÉGRESSION MODIFIÉE

4.1 Aperçu de la régression modifiée

La méthode de régression modifiée est une autre façon

de fournir des estimations composites que l'on peut obtenir comme agrégats pondérés du fichier de données d'enquête courant. La méthode s'applique à un ensemble d'éléments clés déterminé à l'avance, pour lequel elle réalise un nombre particulièrement faible d'erreurs d'échantillonnage.

La technique de régression modifiée fait appel à une régression généralisée pour le fichier de données du mois courant après avoir annexé de nouvelles variables auxiliaires z_H^i à chaque unité i à un moment t . Ici z_H^i est un vecteur de ligne comportant un élément pour chacun des éléments clés. Pour ceux-ci, on a des « pseudo-pères » Z_t^i fondés sur les estimations du mois précédent pour les éléments clés. L'estimateur de régression modifiée est alors donné par une étape de régression généralisée qui applique les repères démographiques aussi bien que les pseudo-pères.

$$y_M^i = y_H^i + (X_t^i, Z_t^i) - (x_H^i, z_H^i) \beta_M^i \quad (7)$$

pour

$$\beta_M^i = \left(\sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' (x_H^i, z_H^i)^{-1} \sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' y_H^i \right)^{-1} \quad (8)$$

$$\text{donc, } y_M^i = \sum_{i \in D} w_{\pi}^n y_H^i \text{ pour}$$

$$w_M^i = w_{\pi}^i \left\{ 1 + (X_t^i, Z_t^i) - (x_H^i, z_H^i) \right\}$$

$$\left(\sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' (x_H^i, z_H^i)^{-1} \sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' y_H^i \right)^{-1} \quad (9)$$

La clé de la méthode est la définition des variables auxiliaires. Soit D , l'ensemble des unités des groupes de renouvellement (ceux comportant des logements sélectionnés aux deux moments) à un moment t . Soit y_H^i , le vecteur d'éléments clés pour l'unité i à un moment t et X_t^i , les totaux de population correspondants. Pour $i \in D$, notons y_{t-1}^i , la valeur du mois précédent pour le vecteur d'éléments clés ou, si aucune valeur n'a été déclarée, imputons $y_{t-1}^i = y_{t-1}^i$ comme l'a suggéré Singh (1996).

J'examine des estimations de régression modifiée pour z_H^i du type ci-dessous, pour $a \in [0, 1]$:

Malheureusement, cela entraîne également une possibilité de biais lorsque des personnes qui ne sont pas représentées dans l'échantillon apparte manifestent un comportement différent de celui des personnes de deux mois.

Le mouvement $y_{t-1}^{HD} - y_{t-1}^{HD}$ en (11) se fonde en réalité uniquement sur l'échantillon apparte (c'est-à-dire les unités qui déclarent aux deux moments), puisque les autres unités des groupes de renouvellement apparte D contribuent zéro au mouvement (pour l'imputation utilisée ici). Il peut en résulter des estimateurs de régression modifiée ayant une erreur d'échantillonnage moins élevée qu'un estimateur AK , car ce « mouvement de l'échantillon apparte » n'est pas influencé par les unités qui ne sont pas présentes pour les deux mois.

4.2 Propriétés des estimateurs de régression modifiée

À noter que $Z_t^i \approx y_{t-1}^{HD}$ puisque $x_{t-1}^i = X_{t-1}^i$. Cela termine la définition des estimateurs de régression modifiée.

$$\sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' (x_H^i, z_H^i)^{-1} \sum_{i \in D} w_{\pi}^n (x_H^i, z_H^i)' y_H^i \quad (13)$$

$$Z_t^i = y_{t-1}^{HD} + (X_t^i - x_{t-1}^i) \beta_{ad}^i \quad (12)$$

donne les repères du mois courant :

régression modifiée du mois précédent pour que l'addition des repères généralisée afin de rajuster l'estimateur de rajustée de façon à correspondre aux poids du mois courant. Sivant Singh et coll. (1997), j'ai utilisé une étape de

Un pseudo-repère Z_t^i approprié serait une estimation de X_{t-1}^i suggérée par Fuller (1999).

(1997). L'utilisation d'une valeur a intermédiaire a été les méthodes MRI et MR2 respectivement de Singh et coll. apparte y_{t-1}^{HD} . Les valeurs $a = 0$ et $a = 1$ donnent mouvement fondé sur les groupes de renouvellement Thompson du mois courant moins une estimation de Horvitz y_{t-1}^{HD} . Pour $a = 1$, z_H^i est l'estimation de Horvitz du mois courant. Pour $a = 0$, z_H^i est simplement l'estimation d'unités en D seulement et des poids de sélection sont des estimations de X_{t-1}^i et de X_t^i respectivement en $y_{t-1}^{HD} = 8/7 \sum_{i \in D} w_{\pi}^n y_{t-1}^i$ et $y_{t-1}^{HD} = 8/7 \sum_{i \in D} w_{\pi}^n y_{t-1}^i$.

$$z_H^i = (1-a)y_{t-1}^{HD} + a(y_{t-1}^i - (y_{t-1}^{HD} - y_{t-1}^{HD})) \quad (11)$$

Compte tenu de cette définition, nous avons

$$= a y_{t-1}^i \quad \text{pour } i \notin D. \quad (10)$$

pour $i \in D$

$$z_H^i = (1-a) \left(\frac{7}{8} y_{t-1}^i + a \right) \left(y_{t-1}^i - \frac{7}{8} y_{t-1}^i - y_{t-1}^i \right)$$

2.4 Biais d'accoutumance

Idéalement, les estimations de groupe de renouvellement devraient comporter la même espérance Y_i , mais dans la pratique elles comportent des espérances légèrement différentes et donc un biais différent. La différence est due en partie à des pratiques de collecte; par exemple, les logements échantillonnés pour la première fois sont interviewés dans le cadre d'une visite sur place, tandis que d'autres groupes de renouvellement sont surtout interviewés par téléphone. Il n'est pas clair quel groupe de renouvellement est influencé le moins par ce genre de biais «d'accoutumance». L'estimation globale comporte un biais d'accoutumance qui est un mélange des biais de chaque groupe de renouvellement. Nous avons recours à de bonnes techniques d'enquête pour réduire ce biais au minimum. À noter que tous les estimateurs composites recevront un apport différent des groupes de renouvellement, et comporteront donc différents biais d'accoutumance.

3. ESTIMATION COMPOSITE AK

3.1 Estimateur composite AK

L'estimateur composite AK (Gurney et Daly 1965) est conçu de façon à insister davantage sur le mouvement des groupes de renouvellement appartenés (c'est-à-dire les groupes de renouvellement dans lesquels les mêmes logements ont été sélectionnés pour le mois courant et le mois précédent). L'estimateur compte trois composantes. La première est une moyenne des estimations de groupe de renouvellement pour les données du mois courant (moment t). La deuxième est l'estimateur composite AK du mois précédent, plus une estimation de mouvement fondée uniquement sur les groupes de renouvellement appartenés. La troisième composante est la différence entre les estimations du groupe de renouvellement non appartené et des groupes appartenés. La part de chaque composante qu'il convient d'utiliser est donnée par deux paramètres A et K, comme suit :

$$y_{AK}^t = (1 - K) \frac{1}{8} \sum_{r=1}^8 y_{Rr}^t + K \left(y_{AK}^{t-1} + \frac{1}{8} \sum_{r=2}^8 y_{Rr}^t - \frac{1}{7} \sum_{r=1}^7 y_{Rr}^{t-1} \right) + A \left(y_{R1}^t - \frac{1}{8} \sum_{r=2}^8 y_{Rr}^t \right) \quad (6)$$

3.2 Choix des valeurs paramétriques

Le paramètre clé est K, qui indique la part de l'estimation nouvelle qui se fonde sur le mouvement des groupes de renouvellement appartenés. La valeur optimale de A et de K à utiliser dépend de la variable qu'il s'agit d'estimer. Des valeurs K plus élevées se prêtent davantage à l'emploi

qu'au chômage, puisque l'emploi comporte une plus forte corrélation d'un mois à l'autre. L'addition des estimations composites AK des personnes ayant un emploi, sans emploi et « ne relevant pas de la population active » ne donne pas exactement la population totale à moins que l'on n'utilise les mêmes paramètres pour toutes les estimations. On choisit donc un compromis pour A et K. Les résultats présentés ici se fondent sur $A = 0,06$ et $K = 0,7$. On a trouvé ces valeurs en faisant l'essai d'une gamme de valeurs de A et de K et en choisissant les valeurs qui donnent des estimations optimales des personnes ayant un emploi. Dans la présente étude, aucune valeur de A et de K n'a donné des estimations des personnes sans emploi appréciablement meilleures que ces valeurs.

Notre étude empirique n'a pas manifesté un niveau particulièrement bon d'erreurs d'échantillonnage pour l'estimateur AK. Cela est peut-être attribuable au calage fin qui a servi à obtenir les estimations de groupe de renouvellement; il est possible que le recours à des catégories plus vastes améliore le bilan des erreurs d'échantillonnage.

3.3 Propriétés de l'estimateur AK

L'estimateur AK insiste davantage sur le mouvement au sein des groupes de renouvellement appartenés. Ainsi, le groupe de renouvellement qui contient des logements admissibles pour la première fois contribue moins que dans l'estimateur RG. L'estimateur AK comporte donc un biais d'accoutumance différent de celui de l'estimateur RG.

L'estimateur AK est récursif : il faut l'estimateur du mois précédent pour préparer celui du mois courant. Cela est peu pratique lorsqu'il s'agit de préparer des estimations pour un nouvel élément ou une nouvelle catégorie. De plus, le besoin d'utiliser les mêmes valeurs de A et de K pour tous les éléments peut entraîner des estimations sous-optimales pour un élément donné.

À cause de ces préoccupations, la Current Population Survey des États-Unis a été changée en fonction d'une variante appelée « pondération composite AK » (Lent, Miller et Cantirell 1994). Dans la pondération composite AK, on prépare des estimations distinctes de l'emploi et du chômage pour un certain nombre de catégories publiées importantes, à l'aide de la stratégie composite AK avec des paramètres optimaux pour l'estimation en question. Les données courantes sont alors calées de façon que l'addition des poids unitaires donne ces estimations AK de même que des repères démographiques. Toutes les estimations sont alors préparées à partir du fichier de données courant à l'aide de ces nouveaux « poids composites AK ».

Le fait de pouvoir préparer les estimations sous forme de somme pondérée des données d'un même mois est un grand avantage de la stratégie de pondération composite AK. Un autre avantage est que les plus importantes estimations sont des estimations composites AK avec un choix de AK presque optimal. Un inconvénient est que seules les plus importantes estimations sont de véritables estimations composites. Toutes les autres estimations (y

2. ESTIMATEURS COURANTS POUR L'ENQUÊTE SUR LA POPULATION ACTIVE

2.1 Aperçu de l'EPA

L'EPA comporte un plan d'échantillonnage à plusieurs

degrés, le premier étant un échantillon de petites régions géographiques appelées CD (« Census collector's Districts ». On tire un nouvel échantillon de CD tous les cinq ans, et les CD sont classés dans huit « groupes de renouvellement ». Les logements tirés d'un CD restent dans l'échantillon durant huit enquêtes, puis ils sont remplacés par d'autres logements tirés du même CD. On appelle renouvellement ce remplacement des logements, tous les logements d'un groupe de renouvellement étant remplacés en même temps. Des intervieweurs cherchent à recueillir des données pour toutes les personnes admissibles dans les logements sélectionnés.

Dans l'EPA, la situation d'activité de la personne (ayant un emploi, sans emploi, ne relevant pas de la population active) revêt un intérêt particulier. Le nombre de personnes se rapportant à chaque situation d'activité, pour diverses catégories de personnes, est un élément clé qu'il s'agit d'estimer dans l'enquête. Les nombreux utilisateurs des données d'enquête accordent une importance encore plus grande aux estimations du mouvement des chiffres d'un moment au suivant. On peut affirmer que les indications à plus long terme de la direction de la série sont encore plus importantes (par exemple le mouvement de la tendance X_{t-1} ou d'une tendance semblable plus lisse (Bell 1999)).

Le plan d'échantillonnage garantit que la probabilité inconditionnelle de sélection π_{it} est connue pour chaque personne échantillonnée à un moment t . Cela permet d'obtenir un estimateur simple pour un total de population dû à Horvitz et Thompson (1952). Si Y_{it} est l'élément de population qu'il s'agit d'estimer à un moment t , et si Y_{it}^* est le même élément signalé par la i^{e} unité à un moment t , l'estimateur de Horvitz-Thompson est

$$(1) \quad \hat{Y}_H = \sum_i^I w_{it}^* Y_{it}^*$$

pour $w_{it}^* = \pi_{it}^{-1}$, connus comme les poids de sélection.

2.2 L'estimateur de régression généralisée (RG)

La régression généralisée est une méthode de rajustement ou de « calage » d'une série de poids unitaires dont l'addition donne une série d'attributs de la population appelés repères. Pour un choix approprié de repères, les poids résultants donnent une meilleure estimation en tenant compte de renseignements externes.

Dans l'EPA, nous commençons par un calage des poids de Horvitz-Thompson de façon que leur addition donne des repères démographiques indiquant le nombre de personnes dans la population pour 560 strates après sélection (14 régions géographiques classées selon le sexe et 20 groupes d'âge). Les poids d'une strate après sélection particulière sont répartis proportionnellement de façon que

leur addition donne le repère de strate. Cet estimateur par quotient stratifié après sélection est un cas particulier de l'estimateur de régression généralisée ou estimateur RG. Soit x_{it}^* un vecteur de ligne de variables auxiliaires pour l'unité i à un moment t , et $x_{it} = \sum_j b_{jt}^* x_{it}^*$ des estimations du vecteur de ligne correspondant de valeurs repères X_{it} , d'après certains poids initiaux b_{jt}^* . L'estimateur RG fondé sur ces poids initiaux est alors donné par

$$(2) \quad \hat{y}_{it}^* = y_{it}^* + (X_{it}^* - \hat{x}_{it}^*) \beta$$

$$(3) \quad \text{pour } \beta = \left(\sum_i^I b_{it}^* x_{it}^* x_{it}^{*'} \right)^{-1} \sum_i^I b_{it}^* x_{it}^* y_{it}^*$$

$$\text{Donc, } \hat{y}_{it}^* = \sum_i^I w_{it}^* y_{it}^* \text{ pour}$$

$$(4) \quad w_{it}^* = b_{it}^* \left(1 + (X_{it}^* - \hat{x}_{it}^*) \left(\sum_i^I b_{it}^* x_{it}^* x_{it}^{*'} \right)^{-1} \sum_i^I b_{it}^* x_{it}^* y_{it}^* \right)$$

Dans une estimation par quotient stratifiée après sélection, les vecteurs de ligne x_{it}^* contiennent des zéros sauf dans la colonne qui correspond à la strate après sélection de l'unité, et b_{it}^* sont les poids de sélection w_{it}^* . Dans ce cas, les paramètres de régression sont simplement les moyennes de strates après sélection, estimées à l'aide des poids de sélection.

2.3 Estimations par groupe de renouvellement

Chaque groupe de renouvellement est constitué d'un échantillon représentatif de logements, et peut donc fournir une estimation distincte. On numérote les groupes de renouvellement à tel moment selon le nombre de fois que les logements du groupe de renouvellement ont été échantillonnés. On écrit $R(t, i) = r$ si l'unité i se trouve dans le groupe de renouvellement échantillonné pour la r^{e} fois à un moment t . L'estimation de Horvitz-Thompson de Y_{it} d'après le groupe de renouvellement r est

$$(5) \quad \hat{y}_{Hr}^* = \sum_{i: R(t,i)=r}^I 8 w_{it}^* y_{it}^*$$

On peut utiliser la régression généralisée pour améliorer ces estimateurs en calant les poids de façon que leur addition donne une série de repères. Malheureusement, la taille plus faible de l'échantillon dans un groupe de renouvellement unique risque d'exiger l'utilisation d'un plus petit nombre de repères que pour le cas global. Dans la situation de l'EPA, l'auteur du présent exposé a appliqué une même étape de régression généralisée à l'ensemble de l'échantillon de façon que, pour tout l'échantillon, l'addition des poids donne les repères pour les 540 strates après sélection courantes, tandis que, dans chaque groupe de renouvellement, l'addition des poids donne un huitième des repères pour 71 strates après sélection groupées. Les poids résultants, lorsqu'ils sont appliqués à un groupe de renouvellements donne r et multipliés par huit, donnent les estimations de groupe de renouvellement \hat{y}_{it}^* .

Comparaison d'autres estimateurs pour l'Enquête sur la population active

PHILIP BELL¹

RÉSUMÉ

L'auteur examine un choix d'estimateurs applicables à une enquête auprès des ménages périodique comportant un chevauchement contrôlé entre les enquêtes successives. Le thème commun des estimateurs est le recours à des données de moments antérieurs afin d'améliorer les estimations courantes, en bénéficiant de corrélations dans l'échantillon chevauchant. L'auteur considère tous les estimateurs de ce genre comme des estimateurs composites.

Les estimateurs sont évalués en fonction de l'enquête sur la population active (EPA) de l'Australie. Dans l'EPA, on contrôle le chevauchement en divisant l'échantillon du premier degré (régions géographiques) en huit « groupes de renouvellement » parmi lesquels on sélectionne des logements. Chaque mois, les mêmes logements sont tirés de sept des groupes de renouvellement, de nouveaux logements étant sélectionnés à même le groupe restant. L'échantillon est constitué de personnes âgées de 15 ans ou plus demeurant dans les logements sélectionnés.

Ce plan d'échantillonnage entraîne un chevauchement élevé de l'échantillon entre deux mois consécutifs au sein des sept « groupes de renouvellement » appariés. Le fait d'utiliser uniquement des données de ces groupes de renouvellement plutôt que de l'échantillon tout entier permet de diminuer l'erreur d'échantillonnage pour une estimation du mouvement d'un mois à l'autre. À l'aide de techniques d'estimation composite, on peut exploiter cette situation de façon à préparer des estimations comportant une erreur d'échantillonnage plus faible.

À la section 2, l'auteur présente l'EPA de l'Australie et son estimateur courant de « régression généralisée ». L'auteur aborde également la question du biais d'accoutumance (appelé biais de groupe de renouvellement par Bailar 1975).

À la section 3, l'auteur présente l'estimateur « composite AK » proposé par Gurney et Daly (1965). Cette méthode est

MOTS CLÉS : Estimateur composite; meilleur estimateur linéaire sans biais; régression modifiée; enquêtes répétées.

1. INTRODUCTION

L'auteur examine un choix d'estimateurs applicables à une enquête-ménage périodique comportant un chevauchement contrôlé entre les enquêtes successives. Le thème commun des estimateurs est le recours à des données de moments antérieurs afin d'améliorer les estimations courantes, en bénéficiant de corrélations dans l'échantillon chevauchant. L'auteur considère tous les estimateurs de ce genre comme des estimateurs composites.

Les estimateurs sont évalués en fonction de l'enquête sur la population active (EPA) de l'Australie. Dans l'EPA, on contrôle le chevauchement en divisant l'échantillon du premier degré (régions géographiques) en huit « groupes de renouvellement » parmi lesquels on sélectionne des logements. Chaque mois, les mêmes logements sont tirés de sept des groupes de renouvellement, de nouveaux logements étant sélectionnés à même le groupe restant. L'échantillon est constitué de personnes âgées de 15 ans ou plus demeurant dans les logements sélectionnés.

Ce plan d'échantillonnage entraîne un chevauchement élevé de l'échantillon entre deux mois consécutifs au sein des sept « groupes de renouvellement » appariés. Le fait d'utiliser uniquement des données de ces groupes de renouvellement plutôt que de l'échantillon tout entier permet de diminuer l'erreur d'échantillonnage pour une estimation du mouvement d'un mois à l'autre. À l'aide de techniques d'estimation composite, on peut exploiter cette situation de façon à préparer des estimations comportant une erreur d'échantillonnage plus faible.

À la section 2, l'auteur présente l'EPA de l'Australie et son estimateur courant de « régression généralisée ». L'auteur aborde également la question du biais d'accoutumance (appelé biais de groupe de renouvellement par Bailar 1975).

À la section 3, l'auteur présente l'estimateur « composite AK » proposé par Gurney et Daly (1965). Cette méthode est

utilisée dans la Current Population Survey des États-Unis depuis de nombreuses années. Un prolongement appelé « pondération composite AK » est utilisé depuis quelques années; il a été proposé par Fuller (1990) et étudié par Lent, Miller et Camwell (1994, 1996).

À la section 4, l'auteur présente la méthode « de régression modifiée » de l'estimation composite (Singh et Merikou 1995; Singh 1996). Dans le présent exposé, l'auteur examine l'estimateur MR2 de Singh et coll. (1997), qui permet la plus forte réduction de l'erreur d'échantillonnage. L'auteur présente également une variante de cette méthode suggérée par Fuller (1999) en vue de l'Enquête sur la population active du Canada.

À la section 5, l'auteur présente un « meilleur estimateur linéaire sans biais » (MELSB) fondé sur des données tirées d'une « fenêtre » comportant un nombre fixe de mois (1942) dans le cas de 2 occasions. Un MELSB fondé sur toutes les occasions d'une longue série semble peu pratique, bien que Yansaneh et Fuller (1998) en aient élaboré une approximation récurrente. L'auteur améliore le MELSB à fenêtre fixe décrit par Bell (1998) à l'aide de la technique de régression généralisée.

À la section 6, on trouve les résultats de l'application des différentes méthodes à l'estimation de personnes ayant un emploi et des personnes sans emploi dans l'EPA. Les erreurs-types sont estimées pour des indicateurs à plus long terme comme la tendance et le mouvement de la tendance, de même que pour des estimations du niveau mensuel et de son mouvement. L'auteur examine les biais éventuels, même que les indications d'un changement de profil saisonnier.

L'auteur termine l'exposé en comparant les avantages et le inconvénients des différents types d'estimateurs en fonction de l'EPA. L'estimateur MELSB amélioré est jugé efficace et, lorsqu'il est appliqué à l'EPA, il n'est pas exposé à un biais appréciable.

¹ Philip Bell, Bureau of Statistics de l'Australie, courriel : philip.bell@abs.gov.au

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.

HANSEN, M.H., HURWITZ, W.N., NISSELSOHN H. et STEINBERG, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50, 701-719.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, B*, 42, 221-226.

KUMAR, S., et LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Techniques d'enquête*, 9, 1-28.

LENT, J., MILLER, S.M., CANTWELL, P.J. et DUFF, M. (1999). Effect of composite weights on some estimates from the Current Population Survey. *Journal of Official Statistics*, 14, 431-448.

PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, B*, 12, 241-255.

PFEEFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SCOTT, A.J., SMITH, T.M.F. et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.

SINGH, M.P., DREW, J.D., GAMBINO, J. et MAYDA, F. (1990). *Méthodologie de l'Enquête sur la population active du Canada*. Numéro de catalogue 71-526, Statistique Canada.

TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

YANSANEH, I.S., et FULLER, W.A. (1998). Méthode optimale d'estimation réursive pour les enquêtes répétitives. *Techniques d'enquête*, 24, 33-42.

respectivement, tandis que la variance de niveau et la variance de changement pour $\alpha = 0,75$ sont de 0,67 et de 0,13 environ, respectivement, exprimées en unités communes.

La valeur α plus petite offre l'avantage que l'estimateur composite est plus proche de l'estimateur direct. Ainsi, le biais éventuel associé à l'estimateur composite devrait être plus petit pour une valeur α plus petite.

Tableau 3
Efficacités approximatives d'estimateurs compromis
relativement à $\alpha = 1$

p	b_0	$1 - \lambda_n$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$	$\hat{\mu}_t$	$\hat{\mu}_t - \hat{\mu}_{t-1}$
0,70	0,625	0,0625	1,052	0,999	1,069	0,995
0,80	0,741	0,0432	1,099	0,994	1,129	0,984
0,90	0,865	0,0224	1,238	0,975	1,303	0,946
0,95	0,931	0,0114	1,502	0,936	1,616	0,875
0,98	0,972	0,0046	2,177	0,833	2,321	0,712

6. PROBLÈME DE DÉRIVE

Comme il a été noté dans l'introduction, l'estimateur MR2 pourrait s'écarter de l'estimateur direct de façon appréciable, et cet écart pourrait s'étendre sur une longue période. Nous allons maintenant expliquer ce phénomène. Nous pouvons exprimer l'écart de l'estimateur de régression compromis $\hat{\mu}_t$ en fonction de x_{3nt} , vis-à-vis de la moyenne réelle μ_t sous la forme

$$\hat{\mu}_t - \mu_t = (\lambda_0 p)^t (\hat{\mu}_0 - \mu_0) + \sum_{j=0}^{t-1} (\lambda_0 p)^j [\lambda \hat{\mu}_{m,j-1} + (1 - \lambda) (\hat{y}_{B,j-1} - \mu_{j-1})], \quad (6.1)$$

où μ_0 est la moyenne au début du processus et

$$\hat{\mu}_{m,t} = \hat{y}_{m,t} - \mu_t - p (\hat{y}_{m,t-1} - \mu_{t-1}).$$

Si p est proche de un et si nous utilisons $\lambda = 1$, l'erreur $\hat{\mu}_t - \mu_t$ se comporte plus ou moins comme une marche aléatoire pouvant entraîner de longues périodes où $\hat{\mu}_t - \mu_t$ porte le même signe. Par contre, si $\alpha < 1$ et si $p = 1$, $\lambda < 1$ et l'erreur $\hat{\mu}_t - \mu_t$ manifeste moins de dérive. Par exemple, si $\alpha = 0,70$, la corrélation entre des erreurs adjacentes $\hat{\mu}_t - \mu_t$ ne sera pas supérieure à 0,95 suivant l'hypothèse (3,2)-(3,5). Pour l'estimateur MR2, $\lambda = 1$ à mesure que $p \rightarrow 1$ et, par conséquent, l'estimateur MR2 peut manifester une dérive pour p proche de un.

7. CONCLUSION

Par souci de simplicité, nous avons souvent supposé un échantillonnage aléatoire simple afin d'obtenir des résultats théoriques. Des résultats semblables sont valables pour des plans complexes et des variables auxiliaires

supplémentaires, si l'on considère p comme une autocorrélation partielle. De plus, nous avons utilisé des variables x_3 correspondant à une variable y seulement, mais plusieurs variables y peuvent servir à construire les variables x correspondantes en vue de l'estimation de régression. Gambino, Kennedy et Singh (2001) ont mené une étude empirique avec des données de l'EPA en faisant appel à plusieurs variables x_3 comportant un α commun, et ont obtenu un α compromis en vue de l'EPA.

À la section 2.1, nous avons supposé l'absence de non-réponse afin que l'imputation ne soit pas requise. Dans l'EPA, toutefois, il peut y avoir une non-réponse pour un élément y soit au moment $t - 1$, soit au moment t ou aux deux à la fois. Gambino, Kennedy et Singh (2001) décrivent en détail les méthodes d'imputation utilisées concrètement dans l'EPA.

REMERCIEMENTS

Les recherches de Wayne Fuller ont été appuyées en partie par un accord de coopération (43-3AEU-3-80088), conclu entre l'Iowa State University, le National Agricultural Statistics Service et le Bureau of the Census des États-Unis. Nous remercions Harold Mameit d'avoir lu le manuscrit soigneusement, car ainsi des améliorations ont pu être apportées.

BIBLIOGRAPHIE

BELL, P. (2001). Comparaison d'autres estimateurs pour l'Enquête sur la population active. *Techniques d'enquête*, 27, 57-68.

BELL, W.R. et HILLMER, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.

BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées - Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.

BREAU, P., et ERNST, L. (1983). Alternatives estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3^e édition. New York : John Wiley and Sons.

ECKLER, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26, 664-180.

FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.

GAMBINO, J.G., KENNEDY, B. et SINGH, M.P. (2001). Estimation composite par régression pour l'Enquête sur la population active du Canada : Évaluation et application. *Techniques d'enquête*, 27, 69-79.

LINDEBER, J. (1998). *Méthodologie de l'Enquête sur la population active du Canada*. Numéroteur de catalogue 71-526, Statistique Canada.

5. UN ESTIMATEUR COMPROMIS

D'après le tableau 2, l'efficacité de l'estimateur de changement MR2 pour l'EPA fondé sur x_{1p} pour le cas de non-révision, est assez bonne. L'estimateur de changement sans révision MRI fondé sur x_{2t} offre une efficacité relativement médiocre puisqu'il est membre de la classe (3.7) avec $\lambda = \theta = 0,8333$. Par contre, l'estimateur de niveau MRI fondé sur x_{2t} est supérieur à l'estimateur MR2 fondé sur x_{1p} et il y a des membres de la classe (3.7) qui sont nettement supérieurs à l'estimateur de niveau MR2. Puisque la valeur de λ dans l'estimateur MR2 est relativement grande et puisque la valeur de λ pour l'estimateur MR1 est relativement faible, nous pouvons créer des approximations des membres les plus intéressants de la classe (3.7) comme combinaisons linéaires de (2.10) et de (2.5). Notons

(5.1)
$$x_{3,nt} = \alpha x_{1,nt} + (1 - \alpha)x_{2,nt}$$

où $0 < \alpha \leq 1$ est un nombre fixe. L'estimateur de régression fondé sur $x_{3,nt}$ donne une approximation d'un membre de la classe (3.7) avec

(5.2)
$$\lambda = \alpha \lambda_A + (1 - \alpha) \theta,$$

où λ_A est défini en (2.6). Si $p = 0,95$,

$$\lambda = \alpha (0,9886) + (1 - \alpha) (5/6),$$

pour le schéma de renouvellement de l'EPA avec $\theta = 5/6$ et $b_0 = (7 - 2p)^{-1} 5p$; $\lambda = 0,95$ si $\alpha = 0,75$.

Nous choisissons α pour avoir la combinaison souhaitée de $\hat{y}_{B,t}$ et de l'« estimateur de régression » fondé sur des observations dans l'ensemble A. Si l'on ne révisé pas l'estimateur de μ_{t-1} , la combinaison préférée dépend de l'importance relative accordée à la variance de niveau et à la variance de changement.

Le tableau 3 indique la variance de l'estimateur MR2 ($\alpha = 1$) relativement à la variance de l'estimateur construit à l'aide de $\alpha = 0,75$ et à la variance de l'estimateur construit à l'aide de $\alpha = 0,65$. Une entrée du tableau 3 pour μ_t est l'expression (3.10) évaluée à λ_A de (2.6) et p , divisée par (3.10) évaluée à λ de (5.2) et p . Une entrée pour $\mu_t - \mu_{t-1}$ est l'expression (4.1) évaluée à λ_A de (2.6) et p , divisée par (4.1) évaluée à λ de (5.2) et p . Ce sont là des approximations d'efficacité réelles car on utilise p pour le coefficient de x_j . Le tableau 3 indique clairement que l'estimateur compromis est légèrement inférieur à l'estimateur MR2 pour un changement d'une période, mais qu'il est bien supérieur à l'estimateur MR2 pour le niveau. Par exemple, avec $p = 0,95$ et $\alpha = 0,65$, l'efficacité relative de l'estimateur compromis est de 1,62 pour le niveau et de 0,87 pour un changement d'une période.

Pour de plus grandes valeurs de p , la variance de changement est bien plus petite que la variance de niveau. Ainsi, pour $p = 0,95$, la variance de niveau et la variance de changement pour $\alpha = 1,00$ sont de 1,00 et de 0,12 environ,

l'habitude de réviser l'estimateur de μ_{t-1} . En l'absence de révision, l'estimateur de changement est $\mu_t - \mu_{t-1}$. En l'absence de révision de μ_{t-1} et compte tenu des conditions (3.2) à (3.5), la variance de $\mu_t - \mu_{t-1}$, où μ_t est défini en (3.7), est la suivante

(4.1)
$$\begin{aligned} V\{\mu_t - \mu_{t-1}\} &= V\{\lambda[\bar{y}_t] + (\mu_{t-1} - \bar{x}_{2,t})p\} \\ &+ (1 - \lambda) \bar{y}_{B,t} - \mu_{t-1} \\ &= [\sigma^{-1} \theta^{-1} \lambda^2 (1 - p^2) + (1 - \lambda)^2 \\ &+ (p\lambda - 1)^2 q_{t-1}^{-1} V\{\bar{y}_{B,t}\}]. \end{aligned}$$

Le tableau 2 contient des variances extrêmes normalisées du changement estimatif, $\mu_t - \mu_{t-1}$, pour un choix de valeurs de g et de λ , avec $g\theta = 5$. Les entrées du tableau sont les variances extrêmes du changement estimatif divisées par la variance du changement d'après les éléments communs, $V\{\bar{y}_{m,t} - \bar{y}_{m,t-1}\}$. La variance du changement fondée sur les éléments communs est de $2\theta^{-1} (1 - \theta)(1 - p) V\{\bar{y}_{B,t}\}$.

Tableau 2
Variances extrêmes normalisées en cas de changement d'une période sans révision : Schéma de renouvellement de l'EPA

λ	p	0,98	0,90	0,95	0,994	0,996
0,833	1,039	1,168	1,550	2,312	4,595	4,996
0,840	1,024	1,142	1,492	2,189	4,277	
0,860	0,989	1,079	1,345	1,872	3,454	
0,880	0,963	1,029	1,223	1,607	2,756	
0,900	0,947	0,993	1,127	1,391	2,181	
0,920	0,940	0,970	1,055	1,222	1,723	
0,940	0,942	0,959	1,007	1,100	1,379	
0,960	0,953	0,961	0,982	1,024	1,146	
0,980	0,972	0,975	0,980	0,991	1,021	
0,990	0,985	0,986	0,987	0,990	0,998	
0,995	0,992	0,993	0,993	0,994	0,996	

Les tableaux 1 et 2 indiquent clairement le risque de ne pas réviser l'estimation de μ_{t-1} . Par exemple, si $p = 0,95$, la variance d'un changement sans révision d'une période est réduite au minimum avec $\lambda = 0,99$, mais la variance de niveau est réduite au minimum avec $\lambda = 0,91$. Une valeur compromise de $\lambda = 0,95$ donne une variance de niveau qui est supérieure de 14 % environ à la valeur optimale et une variance de changement qui est supérieure de 7 % environ à la plus petite variance du tableau 2. Si nous utilisons les valeurs de λ_j associées à l'estimateur MR2, les entrées du tableau 2 sont de 0,940, 0,960, 0,979, 0,989 et 0,996 pour $p = 0,70, 0,80, 0,90, 0,95$ et 0,98, respectivement. Ainsi, en l'absence de révision, et si l'on laisse de côté la différence entre b_0 et p , l'estimateur MR2 est presque optimal comme estimateur de changement, contrairement à l'estimateur MR1, où l'estimateur MR1 correspond à $\lambda = 0,833$ du tableau 2.

où $0 < \lambda \leq 1$ est à déterminer. Afin de réduire la variance du niveau courant au minimum, étant donné μ_{i-1} avec une

variance $q_{i-1}^2 V\{\bar{y}_{B,i}\}$, on réduit

$$V\{\hat{\mu}_i\} = V\{\lambda \hat{\mu}_{m,i} + (1 - \lambda) \bar{y}_{B,i}\}$$

$$= \lambda^2 V\{\hat{\mu}_{m,i}\} + (1 - \lambda)^2 V\{\bar{y}_{B,i}\}, \quad (3.8)$$

au minimum relativement à λ . Suivant les hypothèses (3.3), (3.4) et (3.5), la valeur optimale de λ pour le niveau courant

est

$$\lambda^{\text{opt}} = [\varepsilon^{-1} \theta^{-1} (1 - p^2) + q_{i-1}^2 p^2 + 1]^{-1}.$$

Toutefois, si l'on compte utiliser l'estimateur longtemp, il faut bien comprendre que seules certaines valeurs de q_i sont possibles à la longue. La valeur de λ choisie pour l'estimation de μ_i détermine la variance de $\hat{\mu}_i$ et, par conséquent, elle détermine la variance qui sera intégrée à l'estimateur de μ_{i+1} . Si nous supposons que $\beta = p$, nous

avons

$$V\{\hat{\mu}_i\} = \{\varepsilon^{-1} \theta^{-1} \lambda^2 (1 - p^2) + q_{i-1}^2 p^2 + (1 - \lambda)^2\} V\{\bar{y}_{B,i}\}$$

ou

$$q_{i+1}^2 = \varepsilon^{-1} \theta^{-1} \lambda^2 (1 - p^2) + (1 - \lambda)^2 + \lambda^2 p^2 q_{i-1}^2. \quad (3.9)$$

Ainsi, pour une valeur donnée de λ , la valeur extrême pour q_{i-1}^2

est

$$\lim_{i \rightarrow \infty} q_{i-1}^2 = (1 - \lambda^2 p^2)^{-1} [\varepsilon^{-1} \theta^{-1} \lambda^2 (1 - p^2) + (1 - \lambda)^2]. \quad (3.10)$$

Ce résultat correspond à celui donné par Cochran (1977), page 352, équation (12.86).

Le tableau 1 contient des valeurs des variances extrêmes à mesure que le nombre de périodes devient grand, pour un

choix de valeurs de p et de λ , où $\theta = 5/6$ et $g\theta = 5$ pour l'EPA. Les variances sont normalisées de façon que la

variance de l'estimateur direct fondé sur la moyenne de six panels soit de 1,00. Ainsi, les entrées sont de six fois la

la valeur de λ est posée égale à 0,96, la variance à long terme du niveau courant est de 70 % de celle de l'estimateur direct. Si la valeur de λ est posée égale à 0,90, la variance à long terme est de 58 % de celle de l'estimateur direct

lorsque $p = 0,95$.

La première ligne du tableau 1 est réservée pour $\lambda = 5/6 = \theta$. C'est là la valeur de λ qui correspond à l'utilisation de x_{2i}^* dans un estimateur de régression. La

variance comportant $\lambda = 5/6$ est toujours plus faible que celle de l'estimateur direct à cause de l'amélioration associée à l'utilisation de l'estimateur de régression $\hat{\mu}_{m,i}$. Si $p \neq 0$, l'estimateur de régression comportant x_{2i}^* entraîne une réduction significative de la variance relativement à

4. VARIANCE D'UN CHANGEMENT D'UNE PÉRIODE

Étant donné $\hat{\mu}_{i-1}^*$, $\bar{y}_{m,i-1}^*$, $\bar{y}_{m,i}^*$ et $\bar{y}_{B,i}^*$, l'estimateur optimal de μ_{i-1} n'est plus $\hat{\mu}_{i-1}^*$, parce que $\bar{y}_{m,i}^*$ comporte des renseignements au sujet de μ_{i-1} . Toutefois, on n'a pas

celle de \bar{y}_i .
pour $\lambda = 0,833$ dans le tableau 1 et toujours supérieure à l'estimateur direct, \bar{y}_i . L'efficacité de l'estimateur MRL est celle niveau qui est essentiellement la même que pour l'estimateur MRL2 pour le niveau courant offre une efficacité de 0,70, 0,80, 0,90, 0,95 et 0,98, respectivement. Ainsi, l'estimateur λ_i sont de 0,986, 0,982, 0,978, 0,976 et 0,975, pour $p = 0,70, 0,80, 0,90, 0,95$ et 0,98, respectivement. D'après le tableau 1, les variances normalisées de $\hat{\mu}_i$ pour ces valeurs lors, $\lambda_i = (1 - \theta) (1 - b_0)$, où b_0 est donné par (2.4). Dès si nous utilisons la valeur extrême b_0 de b , nous avons $b^* = 0,9422$. Si $p = 0,90$, on a $b_0^* = 0,8659$ et $b^* = 0,8853$. la classe. Par exemple, si $p = 0,95$, on a $b_0^* = 0,9314$ et où b^* est « proche » de p , il est « proche » d'un membre de MRL2 n'est pas un membre de la classe (3.7), dans la mesure où λ_i et b^* sont définis en (2.6). Même si l'estimateur

$$\hat{\mu}_i = \lambda_i [\bar{y}_{m,i}^* + (\hat{\mu}_{i-1}^* - \bar{y}_{m,i-1}^*) b^*] + (1 - \lambda_i) \bar{y}_{B,i}^*,$$

s'écrit sous la forme

Nous considérons maintenant l'estimateur MRL2 (2.3) qui la valeur optimale de λ est de 0,93 environ.

la valeur optimale de λ est de 0,91 environ et pour $p = 0,98$ la valeur optimale de λ est de 0,85 environ, pour $p = 0,95$ la valeur optimale de λ est de 0,833; pour $p = 0,7$ la valeur optimale de λ est une fonction de p et elle

λ	0,70	0,80	0,90	0,95	0,98
0,833	0,897	0,840	0,743	0,665	0,600
0,840	0,895	0,836	0,740	0,650	0,581
0,860	0,894	0,830	0,714	0,614	0,527
0,880	0,903	0,835	0,705	0,588	0,481
0,900	0,921	0,851	0,711	0,575	0,444
0,920	0,951	0,882	0,736	0,582	0,420
0,940	0,992	0,928	0,785	0,617	0,420
0,960	1,046	0,994	0,867	0,698	0,465
0,980	1,115	1,083	0,997	0,861	0,619
0,990	1,155	1,138	1,087	0,998	0,803
0,995	1,177	1,168	1,140	1,089	0,960

Tableau 1
Variances de niveau extrêmes normalisées :
Schéma de renouvellement de l'EPA

L'estimateur direct, \bar{y}_i , qui utilise uniquement des données courantes.

2.2 Un autre estimateur

Il est possible de définir d'autres variables de régression à utiliser dans l'estimation composite de régression. Nous présentons une variable de régression particulière ci-dessous. L'estimateur de régression associé n'est pas proposé comme estimateur ultime, mais l'estimateur est membre d'une classe pour laquelle des calculs d'efficacité sont donnés. Un autre estimateur que l'estimateur MR_2 de Singh (1996) est décrit à la section 5.

$$(2.7) \quad \begin{array}{l} x_{2,i} = \gamma_{i-1,i} \quad \text{si } i \in A_i \\ = q_{i-1} \quad \text{si } i \in B_i \end{array}$$

Si l'on utilise cette variable dans un estimateur de régression, la moyenne de contrôle est $\hat{\mu}_{t-1}$, l'estimateur de la période précédente, puisque la moyenne pour la variable créée sert à estimer la moyenne pour la période $t-1$. Singh (1996) a utilisé une variable x_{2it}^* semblable à x_{2it} . Dans la variable de Singh, le $\hat{\mu}_{t-1}$ en (2.7) est $\hat{y}_{m,t-1}$ si $i \in B_t^*$. Considérons un estimateur de régression construit à l'aide de x_{2it}^* et rappelons que la moyenne de contrôle de x_{2it}^* est $\hat{\mu}_{t-1}$. L'estimateur de régression utilisant x_{2it}^* peut s'écrire

$$(2.8) \quad \hat{u}^{\text{reg}, t}_i = \hat{y}_i + (\hat{u}^{t-1}_{2,i} - \hat{\beta}_i)$$

10

$$\text{Cov}\{\bar{y}_{B,t}, (\bar{y}_{m,t} - \bar{y}_{m,t-1}\beta)\} = 0, \quad (3.5)$$

où g est le nombre de groupes de renouvellement (panels). L'hypothèse (3.1) est raisonnable si les panels originaux comportent une fonction de covariance dont celle d'un processus autorégressif d'ordre un représente une bonne approximation. Pour l'EPA, les covariances nulles en (3.4) et en (3.5) et l'hypothèse (3.2) sont uniquement des approximations puisque \hat{y}_t ne se fonde pas sur un échantillon entièrement indépendant.

Nous écrivons l'estimateur de régression en fonction du chevauchement sous la forme

$$\mathfrak{g}^{l^{-1}}\mathfrak{v} + \mathfrak{g}^{l^{-1}m}\underline{\mathfrak{y}} - {}^{l'm}\underline{\mathfrak{y}} = {}^{l'm}\mathfrak{v}$$

et, suivant les hypothèses, nous obtenons

$$(3.6) \quad \{ \cdot, \cdot \}_{B,1} \wedge [z^d b + (z^d - 1)_{I-} \theta_{I-} \delta] = {}^{(1w)} \eta \wedge$$

Pour l'EPA, $g = 6$ est le nombre de panels. Considérons maintenant un estimateur qui est une combinaison linéaire

$$\hat{p}_{m,t} = \hat{y}_{m,t} + (\hat{p}_{t-1} - \hat{y}_{m,t-1}) \hat{\beta}, \quad (2.10)$$

où β est la régression de y_i sur y_{i-1} dans l'ensemble A , est l'estimateur linéaire optimal pour μ , fondé sur μ_{i-1} et les

pour le calcul de x_{1t}^{it} , et de l'estimateur MR2 résultant. Des variables de contrôle supplémentaires ayant la forme (2.2) associées à d'autres variables y de même que des variables auxiliaires comportant des totaux de population connus sont également comprises dans l'estimation de régression. Les variables auxiliaires de l'EPA englobent des variables démographiques, comme l'âge, le sexe et l'emplacement. Les variables x particulières en (2.2) sont conçues de façon que le total estimé de x_1 soit un estimateur du total de y à la période précédente. Ainsi, le total de contrôle pour x_1 dans la procédure de régression est l'estimateur du total de y à la période précédente.

Afin de simplifier la présentation, nous parlons du modèle d'échantillonnage aléatoire simple sans x_{Ct} . Les résultats s'étendent au cas général si l'on considère le paramètre p comme étant la corrélation partielle entre y_t et y_{t-1} après rajustement pour x_{Ct} . Dans le modèle autorégressif avec p fixe, une coordonnée à l'origine et pas d'autre x_{Ct} dans le modèle, il est possible de montrer que b_t converge pour ce qui est de la probabilité vers

$$b_0 = p \lim_{n \rightarrow \infty} b_t = \theta p [2 - \theta - 2(1 - \theta)p - (1 - \theta)\sigma_y^2 \Delta_2^{-1}]^{-1},$$

où $\Delta_2^2 = (\mu_t - \mu_{t-1})^2$. Si nous supposons que $(1 - \theta)\sigma_y^2 \Delta_2^2$ est faible relativement aux autres termes, nous obtenons

$$b_0 \approx \theta p [2 - \theta - 2(1 - \theta)p]^{-1}. \quad (2.4)$$

Pour l'EPA, $b_0 = (7 - 2p)^{-1} 5p$.

On obtient d'autres représentations pour l'estimateur $\hat{\mu}_t$ en omettant x_{Ct} , grâce à la formule $\hat{y}_t = \theta \hat{y}_{mt} + (1 - \theta)\hat{y}_{Bt}$. Ainsi

$$\hat{\mu}_t = (1 - b_t)\hat{y}_t + [\hat{\mu}_{t-1} + (\hat{y}_{mt} - \hat{y}_{m,t-1})]b_t$$

$$= \theta [\hat{y}_{mt} + (\hat{\mu}_{t-1} - \hat{y}_{m,t-1})b_t] + (1 - \theta)(1 - b_t)\hat{y}_{Bt}$$

$$+ (1 - \theta)(\hat{\mu}_{t-1} - \hat{y}_{m,t-1})\hat{y}_{mt} + \hat{y}_{mt}b_t + (1 - \theta)(1 - b_t)\hat{y}_{Bt}$$

$$= [\theta + (1 - \theta)b_t][\hat{y}_{mt} + (\hat{\mu}_{t-1} - \hat{y}_{m,t-1})b_t] + (1 - \theta)(1 - b_t)\hat{y}_{Bt}$$

$$= \lambda_A^v [\hat{y}_{mt} + (\hat{\mu}_{t-1} - \hat{y}_{m,t-1})b_t] + (1 - \lambda_A^v)\hat{y}_{Bt}, \quad (2.5)$$

où

$$1 - \lambda_A^v \approx (1 - \theta)(1 - b_0)$$

et

$$b \approx [\theta + (1 - \theta)b_0]^{-1} b_0. \quad (2.6)$$

La première expression à la droite de l'égalité de (2.5) donne l'estimateur MR2 comme combinaison linéaire de l'estimateur direct \hat{y}_t et de l'estimateur de différence $\hat{\mu}_{t-1} + (\hat{y}_{mt} - \hat{y}_{m,t-1})$, c'est-à-dire sous la forme d'un estimateur composite. L'expression finale de (2.5) donne l'estimateur comme combinaison linéaire d'un estimateur de type régression « fondé sur les panels chevauchants et la moyenne des premiers panels.

où $\hat{y}_{t,t(i)}^{it}$ est la valeur prévue dans la régression de $y_{t,i}$ sur x_{Ct} et $d_{t,t(i)}^{it}$ est l'écart de la valeur prévue de régression. On a alors

$$x_{1t}^{it} = \hat{y}_{t-1,t(i)-1}^{it} + d_{t-1,t(i)-1}^{it} - \hat{y}_{t,t(i)}^{it} - d_{t,t(i)}^{it}$$

$$+ \hat{y}_{t,t(i)}^{it} + d_{t,t(i)}^{it} \quad \text{si } i \in A_t$$

$$= \hat{y}_{t,t(i)}^{it} + d_{t,t(i)}^{it} \quad \text{si } i \in B_t$$

Pour les variables démographiques X_{Ct}^{it} , il est raisonnable de penser que $\hat{y}_{t-1,t(i)-1}^{it}$ est proche de $\hat{y}_{t,t(i)}^{it}$ et proche de x_{Ct}^{it} . Par conséquent, la partie de x_{1t}^{it} qui est orthogonale à x_{Ct}^{it} est proche de

$$x_{d,1,t}^{it} = \theta^{-1} (d_{t-1,t(i)-1}^{it} - d_{t,t(i)}^{it})$$

$$+ d_{t,t(i)}^{it} \quad \text{si } i \in A_t$$

$$= d_{t,t(i)}^{it} \quad \text{si } i \in B_t$$

On peut écrire

$$y_{t,i}^{it} = \hat{y}_{t,t(i)}^{it} + d_{t,t(i)}^{it}$$

où \hat{x}_{Ct}^{it} est la moyenne de population du vecteur de variables auxiliaires, comme l'âge et le sexe, au moment t , \hat{x}_{Ct}^{it} est la moyenne d'échantillon pondérée des variables auxiliaires et $(\hat{\beta}_{Ct}, b_t)^T$ est le vecteur de coefficients de régression pour la régression de y_t sur (x_{Ct}, x_{1t}) .

$$\hat{\mu}_t = \hat{y}_t + (x_{Ct}^{it} - \hat{x}_{Ct}^{it})[\hat{\beta}_{Ct} + \hat{\mu}_{t-1} - (\hat{y}_{m,t-1} - \hat{y}_{mt} + \hat{y}_t)]b_t, \quad (2.3)$$

de régression comme suit

d'une méthode jackknife d'estimation de la variance. Bell (2000) a comparé plusieurs estimateurs composites à l'aide de données tirées de l'enquête sur la population active de l'Australie.

2. ESTIMATION DE RÉGRESSION COMPOSITE

On utilise deux types d'observations dans l'estimation composite : les observations qui se font uniquement au moment t courant et les observations qui se font tant au moment courant qu'au moment précédent, $t-1$. Parfois, l'information des observations précédentes est comprise dans la ou les estimations pour la ou les périodes précédentes. Soit w_t , le poids d'échantillonnage pour l'observation t au moment t , et A_t , l'ensemble d'éléments comportant des observations pour les deux périodes t et $t-1$, et B_t , l'ensemble d'éléments observé uniquement au cours de la période courante t . Dans un tel contexte initial, t est l'indice pour un même répondant. S'il n'y a pas de non-réponse, l'ensemble A_t pour l'EPA est constitué d'individus dans les cinq panels qui se trouvaient dans l'échantillon au cours de la période précédente, appelés panels de chevauchement. En l'absence de non-réponse, l'ensemble B_t pour l'EPA comporte des individus observés d'abord pendant la période courante, appelés le premier panel. Supposons que

$$\sum_{t \in A_t} w_t + \sum_{t \in B_t} w_t = N_t = \text{total de population estimatif.}$$

Soit θ_t , la fraction de l'échantillon qui chevauche au moment t :

$$\theta_t = N_t^{-1} \sum_{t \in A_t} w_t \quad (2.1)$$

Dans l'EPA, θ_t est de 5/6 environ et presque constante au fil des ans. Par souci de simplicité, l'indice t de A_t , B_t , et θ_t , est souvent omis.

2.1 Estimateur

L'estimateur MR2 de Singh (1996) fait appel à la variable de contrôle

$$x_{t-1} = \theta_t^{-1} (y_{t-1} - y_t) + y_t \quad \text{si } t \in A_t$$

$$= y_t \quad \text{si } t \in B_t, \quad (2.2)$$

dans le programme de régression, où y_t est la valeur d'une caractéristique d'intérêt, y , pour l'élément t au moment t . Compte tenu de la non-réponse dans l'EPA, la proposition originale de Singh faisait appel à l'imputation pour les données manquantes et posait $\theta = 5/6$, après l'imputation de données manquantes. Dans notre discussion initiale, nous utilisons la θ_t définie en (2.1), en supposant l'absence de non-réponse de façon que l'imputation ne soit pas requise. À noter que le « micropartiment » de fichiers de données individuels aux moments $t-1$ et t est nécessaire

Les auteurs mentionnés ci-dessus ont fait appel à la stratégie traditionnelle fondée sur le plan, en supposant que les totaux inconnus à chaque occasion étaient des paramètres fixes. D'autres auteurs (Scott, Smith et Jones 1977; Jones 1980; Binder et Dick 1989; Bell et Hillimer 1990; Tiller 1989 et Pfeffermann 1991) ont élaboré des estimations sous-jacentes répétées en supposant que les valeurs sous-jacentes réelles étaient une réalisation d'une série chronologique.

Statistique Canada a considéré l'estimation composite K et AK pour l'Enquête sur la population active à plusieurs reprises au cours des 25 dernières années (Kumar et Lee 1983), mais n'a pas adopté l'estimation composite. On a plutôt calculé, à l'aide d'un programme de poids de régression, un estimateur de régression généralisée fondé uniquement sur les données du mois courant. Lorsque l'estimation composite a été considérée dans les années 1990, il y avait de fortes pressions pour que l'on élabore une procédure d'estimation composite faisant appel au programme d'estimation existant. Singh (1996) a proposé une méthode ingénieuse appelée régression modifiée (MR) pour surmonter la difficulté. Cette méthode donne lieu à un estimateur composite appelé MR2 qui utilise le programme existant de poids de régression. Singh a suggéré la création de variables x servant de variables de contrôle dans le programme de régression. Grâce aux variables créées et à l'estimateur de la période précédente, on utilise le programme existant de poids de régression pour établir des poids de régression qui définissent l'estimateur pour la période courante. On inclut également des variables de contrôle comportant des totaux de population connus. Une étude empirique de l'estimateur MR2 a permis de cerner plusieurs caractéristiques de la procédure. Tout d'abord, la variance estimative d'un changement d'une période est fortement réduite. Deuxièmement, la variance estimative du niveau est souvent semblable à celle de l'estimateur direct. Troisièmement, l'estimateur de niveau peut s'écarter appréciablement de l'estimateur direct, et cet écart peut se prolonger longtemps.

Dans le présent exposé, les auteurs examinent l'efficacité des estimateurs MR2 théoriquement dans le cadre d'un montage simplifié. Ils utilisent également un estimateur « compromis » qui entraîne des gains appréciables en ce qui concerne l'efficacité tant pour le niveau que pour le changement relatif à l'estimateur qui n'utilise que les données du mois courant. L'estimateur composite tient compte également du problème de « dérive » mentionné ci-dessus, et peut être mis en oeuvre en fonction du programme existant de poids de régression. Gambino, Kennedy et Singh (2000) ont évalué l'efficacité des estimations composites pour les données de l'EPA, à l'aide

Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada

WAYNE A. FULLER et J.N.K. RAO¹

WAYNE A. FULLER et J.N.K. RAO¹

RESUME

L'enquête sur la population active du Canada est une enquête mensuelle menée auprès de ménages sélectionnés en fonction d'un plan stratifié à plusieurs degrés. L'échantillon de ménages est divisé en six panels (groupes de renouvellement). Un panel reste dans l'échantillon pendant six mois consécutifs, pour être ensuite éliminé de l'échantillon. Par le passé, un estimateur de régression généralisé, fondé uniquement sur les données du mois courant, a été mis en oeuvre à l'aide d'un programme de régression. Dans le présent exposé, les auteurs examinent des procédures d'estimation composite de régression faisant appel à des données d'échantillon tirées de périodes précédentes et pouvant être mises en oeuvre dans le cadre d'un programme de poids de régression. Singh (1996) a proposé un estimateur composite appelé MR_2 qu'on peut calculer en ajoutant des variables x au programme courant de poids de régression. L'estimateur de Singh est considéré comme plus efficace que l'estimateur de régression généralisé pour un changement d'un période, mais non pas pour l'appréciable, et cet écart peut persister longtemps. Les auteurs proposent un estimateur « compromis », en utilisant un estimateur de régression généralisé de façon s'écarter de l'estimateur de régression généralisé pour le même nombre de variables x que MR_2 , qui est plus efficace pour le niveau et pour le changement de régression que l'estimateur de régression généralisé fondé uniquement sur les données du mois courant. L'estimateur proposé tient compte également du problème de dérive et s'applique à d'autres enquêtes faisant appel à un échantillonnage par renouvellement.

MOTS CLES : Échantillonnage; groupes de renouvellement; combinaison d'estimateur.

1. INTRODUCTION

L'estimation composite est un concept utilisé dans l'échantillonnage d'enquête pour décrire des estimateurs de période courante utilisant l'information de périodes antérieures d'une enquête périodique comportant un plan de renouvellement. Lorsque certaines unités sont observées au cours de certaines périodes, mais non pas toutes, on peut en profiter pour améliorer les estimations pour toutes les périodes.

Statistique Canada, le Bureau of the Census des États-Unis et d'autres bureaux de la statistique font appel à un plan de renouvellement pour les enquêtes sur la population active. L'Enquête sur la population active courante du Canada (EPA) est une enquête mensuelle menée auprès de quelque 59 000 ménages sélectionnés en fonction d'un plan d'échantillonnage stratifié à plusieurs degrés. L'unité d'échantillonnage ultime est le ménage, et un échantillon de ménages est divisé en six panels (groupes de renouvellement). Un groupe de renouvellement reste dans l'échantillon pendant six mois consécutifs, pour ensuite être éliminé complètement de l'échantillon. Ainsi, les cinq systèmes consécutifs de ménages sont communs pour deux mois de l'échantillon. Singh, Drew, Gambino et Mayda (1990) et Gambino, Singh, Dufour, Kennedy et Lindeyer (1998) ont présenté des descriptions détaillées du plan de l'EPA. Dans la Current Population Survey (CPS) des États-Unis, l'échantillon est constitué de huit groupes de renouvellement. Un groupe de renouvellement reste dans

l'échantillon pendant quatre mois consécutifs, quitte l'échantillon pour les huit mois suivants, puis retourne à l'échantillon pour quatre mois consécutifs. Il est alors éliminé complètement de l'échantillon. Il y a donc un chevauchement de 75 % d'un mois à l'autre dans l'échantillon, et un chevauchement de 50 % d'une année à l'autre dans l'échantillon (Hansen, Hurwitz, Nisselson et Steinberg 1955).

Patterson (1950), suivant les travaux initiaux de Jessen (1942), a décrit le fondement théorique du plan et de l'estimation des enquêtes répétées, à l'aide de procédures des moindres carrés généralisées. Pour la CPS, Hansen et coll. (1955) ont proposé un estimateur plus simple, appelé l'estimateur composite K. Gurney et Daly (1965) ont présenté un estimateur composite K amélioré, appelé l'estimateur composite AK avec deux facteurs de pondération f_1 et K. Brea et Ernst (1983) ont comparé d'autres estimateurs à l'estimateur composite K pour la CPS. Rao et Graham (1964) ont étudié des schémas de remplacement optimal pour l'estimateur composite K. Eckler (1955) et Wolter (1979) ont étudié des schémas de renouvellement à deux niveaux comme celui utilisé dans la Retail Trade Survey des États-Unis. Yansaneh et Fuller (1998) ont étudié l'estimation récurisive optimale pour des enquêtes répétées. Fuller (1990) et Lent, Miller, Cantwell et Duff (1999) ont élaboré la méthode des poids composites pour la CPS. On obtient les poids composites en appliquant la méthode itérative aux poids du plan en fonction de taux de contrôle particuliers qui englobent les taux de population de

LENT, J., MILLER, S. et CANTWELL, P. (1994). Composite weights for the current population survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 867-872.

LENT, J., MILLER, S. et CANTWELL, P. (1996). Effects of composite weights on some estimates from the Current Population Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1, 130-139.

LIANG, K.-Y., et ZEGGER, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.

SÄRNDAHL, C.-E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

SINGH, A.C. (1994). Sampling-design-based estimating functions for finite population totals. Invited paper, *Abstracts of the Statistical Society of Canada, Annual Meeting, Banff, Alberta*, 8-11, page 48.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1, 120-129.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1, 120-129.

SINGH, A.C., et FOLSOM, R.E. Jr. (2000). Bias corrected estimating functions approach for variance estimation adjusted for poststratification. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 610-615.

SINGH, A.C., KENNEDY, B., WU, S. et BRISSEBOIS, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 300-305.

SINGH, A.C., et MERKOURIS, P. (1995). Composite Estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 420-425.

WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

La simple vérification qui suit peut être exécutée pour diagnostiquer l'effet du biais dû à l'imputation des données sur la situation d'emploi du mois précédent pour compenser la non-réponse de certains répondants du mois courant. L'idée fondamentale est de calculer un facteur multiplicatif de redressement du biais applicable à l'estimateur D_{t-1} qui comprend les valeurs imputées. Le facteur est défini comme étant le quotient des deux estimations gr de la caractéristique du mois précédent fondées sur le sous-échantillon apparié. Le dénominateur est une estimation gr pour le mois précédent (comportant des valeurs imputées), tandis que le numérateur est une estimation gr pour le mois précédent (ne comprenant pas de valeur imputée) et ils sont tous deux calculés selon une méthode qui n'est pas tout à fait typique. Pour le numérateur, nous utilisons les répondants de la période $t-1$ avec les réponses qu'ils ont données à la période $t-1$ et, après correction des poids de sondage pour la non-réponse, nous construisons l'estimateur gr avec les variables de contrôle pour la période t . Pour le dénominateur, nous supposons que les sous-ensembles formés par chaque paire de sous-échantillons appariés aux périodes $t-1$ et t (ici l'appariement des sous-échantillons est fait l'un en fonction de l'autre, l'un en avance dans le temps et l'autre en retard) n'ont pas de contrepartie à cause de la non-réponse, donc sont statistiquement remplaçables l'un par l'autre. Puis, nous remplaçons les répondants de la période $t-1$ qui n'ont pas répondu à la période t par les non-répondants à la période $t-1$ qui ont été imputés à la période t , avec les réponses qui leur ont été imputées à la période $t-1$, ainsi que les poids de sondage. Ensuite, nous procédons de nouveau à la correction des poids pour la non-réponse et pour la stratification a posteriori de gr (au moyen des variables de contrôle pour t) pour cet échantillon complet modifié à la période $t-1$. Les poids gr ainsi obtenus sont utilisés pour calculer le dénominateur mentionné plus haut. Nous pouvons maintenant examiner la série chronologique de ce facteur sur plusieurs mois pour diagnostiquer le biais dû à l'imputation. S'il n'est pas jugé proche de 1, alors nous pouvons traiter la moyenne du facteur sur plusieurs mois comme un redressement non aléatoire multiplicatif du biais appliqué à D_{t-1} . En pratique, au lieu de rajuster D_{t-1} , il serait préférable, pour faciliter les calculs, de rajuster la nouvelle variable de contrôle $C_{t-1(c)}$ (de l'équation 3.6) pour la caractéristique correspondante en lui appliquant l'inverse du facteur multiplicatif sus-mentionné. Autrement, on pourrait éviter tout bonnement l'imputation si l'on pouvait modifier le questionnaire pour obtenir les données nécessaires sur la période antérieure telle que nous le proposons dans l'introduction.

Dans leur étude, Lent, Miller et Cantwell (1994, 1996) considèrent l'estimateur composite pondéré ak pour la *Current Population Survey* des États-Unis comme un remplacement possible de l'estimateur ak utilisé à l'heure actuelle si $a=0,2$ et $k=0,4$. Selon notre expérience concernant l'estimateur composite pondéré

ak^* pourrait être une meilleure solution de rechange si l'on considère les gains d'efficacité.

REMERCIEMENTS

Le gros de ces travaux de recherche a été réalisé pendant que les premiers auteurs travaillaient à Statistique Canada. Les auteurs remercient M. Sheridan, J.D. Drew, J. Gambino et particulièrement M.P. Singh pour leurs encouragements et plusieurs discussions fort fructueuses. Ils sont reconnaissants à Jon Rao et Wayne Fuller de leurs commentaires et suggestions. Ils remercient aussi J.M. Levesque, P. Lorenz et, tout spécialement, T. Merkouris (grâce auquel ces travaux ont débuté) pour l'aide qu'ils leur ont apportée pour analyser et interpréter les résultats. Enfin, ils remercient l'examineur et le rédacteur en chef adjoint Harold Mantel pour leurs commentaires constructifs lors de la révision de l'article. Les travaux de recherche du premier auteur ont été financés en partie par une bourse du CRSNG obtenue à l'Université Carleton aux termes d'un professorat de recherche adjoint.

BIBLIOGRAPHIE

BALLAR, B.A. (1975). The effect of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.

BINDER, D.A., et HIDIROGLU, M.A. (1988). Sampling in time. *Handbook of Statistics*, 6 : Sampling, Elsevier Science, NY, 187-211.

CASSEL, C.M., SÄRNDAAL, C.-E. et WREBTMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

COCHRAN, W.G. (1977). *Sampling Techniques*. 3^e édition. John Wiley and Sons.

FULLER, W.A. (1990). Analyse d'enquête à passages répétés. *Techniques d'enquête*, 16, 177-190.

FULLER, W.A. (1999). The Canadian Regression Composite Estimation. Manuscript non-publié.

GODAMBE, V.P., et THOMPSON, M.E. (1989). An extension of quasi-likelihood estimation (avec discussion). *Journal Statistical Planning and Inference*, 22, 137-172.

GURNEY, M., et DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 247-257.

HANSEN, M.H., HURWITZ, W.N. et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. 2, John Wiley and Sons.

KUMAR, S., et LEE, H. (1983). Evaluation de l'application d'estimateurs composites à l'enquête sur la population active du Canada. *Techniques d'enquête*, 9, 196-221.

accès sur la variation et sur le niveau. Notons cependant que le problème de l'incohérence interne mentionné dans l'introduction pourrait se poser si α est lié à y . D'autres caractéristiques intéressantes de cette version sont, d'une part, que l'on peut choisir la valeur de α de façon qu'elle soit strictement non nulle (de façon à éviter les écarts prolongés entre les séries gr et rc) et, d'autre part, que le nombre de variables de contrôle supplémentaires ne double pas lorsque l'on inclut à la fois les prédicteurs axés sur le niveau et ceux axés sur la variation, ce qui permet d'introduire un plus grand nombre de variables de contrôle ainsi qu'un plus grand nombre de degrés de liberté dans l'estimation de la variance.

Maintenant, dans (6.1), α peut être considéré de façon grossière comme le quotient des deux coefficients optimaux a^* et k^* , et le facteur k^* à l'extérieur des crochets de (6.2) peut être remplacé par le coefficient de régression (sous-optimal) $b^{(ica)}$. Donc, $C^{(ica)}$ n'est pas l'équivalent de l'estimateur ak^* optimal, mais on pourrait retenir une certaine optimalité (si l'on faisait en sorte que α soit partiel à y) en fixant la contribution relative des prédicteurs

$$C^{(ak^*)} = F'_t + k^* [(1 - a^*/k^*) (\tilde{C}_{t-1(ak^*)} - \tilde{D}_{t-1}^* + B'_t - F'_t) + (a^*/k^*) (\tilde{C}_{t-1(ak^*)} - \tilde{D}_{t-1}^*)]. \quad (6.2)$$

Figure 1 (a) Emploi Ontario, non-désaisonnalisé

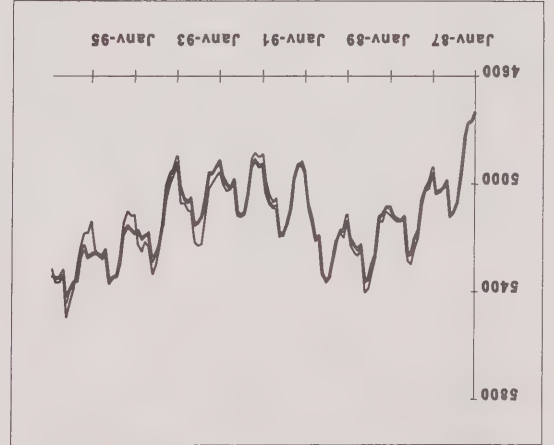


Figure 2 (a) Emploi dans le commerce, Ontario non-désaisonnalisé

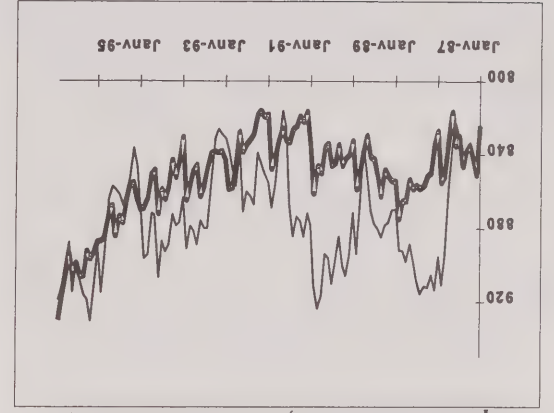


Figure 2 (b)

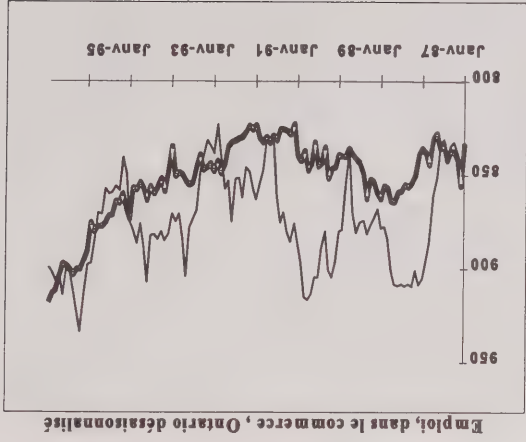
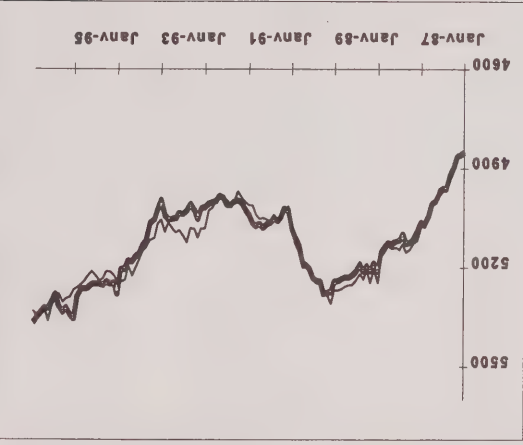


Figure 1 (b) Emploi Ontario, désaisonnalisé



gT
rc
Légende

Tableau 6
Estimations ponctuelles mensuelles au moyen de gr et rc et leurs différences (Ontario, 1996)
(Niveau et variation pour l'emploi dans le secteur du commerce, Ontario, 1996)

Mois	Type	gr	rc	rc-gr
Janvier	Niveau	886,5	858,9	(27,6)
	Variation	-25,8	-21,0	4,8
Février	Niveau	906,5	867,9	38,6
	Variation	20	9,0	-11,0
Mars	Niveau	927,1	874,1	52,9
	Variation	20,6	6,2	-14,4
Avril	Niveau	914,8	872,5	42,3
	Variation	-12,3	-1,6	-10,7
Mai	Niveau	912,8	887,6	25,1
	Variation	-2,1	15,1	-17,2
Juin	Niveau	908,1	888,6	19,5
	Variation	-4,7	0,9	-5,6
Juillet	Niveau	899,9	881,2	18,7
	Variation	-8,2	-7,4	0,8
Août	Niveau	913,9	888,1	25,8
	Variation	14	6,9	-7,1
Septembre	Niveau	886,6	876,4	10,2
	Variation	-27,3	-11,8	-15,6
Octobre	Niveau	888,6	889,3	-0,9
	Variation	12,1	12,9	-0,9
Novembre	Niveau	911,2	902,3	8,9
	Variation	12,6	13,0	-0,4
Décembre	Niveau	917,9	916,3	1,5
	Variation	6,7	14,0	-7,4

Nota : Les écarts-types sont donnés entre parenthèses.

6. CONCLUSIONS

L'estimateur par régression généralisée (gr) utilisé entièrement dans le cadre de l'EPAC produisait des séries instables d'estimations de la variation et de diverses estimations au niveau de domaines. L'estimateur de régression composite (rc) produit des séries d'estimations plus lisses (qui, à leur tour, rendent les estimations de la variation plus stables). La méthode de régression composite s'écarte de l'estimation composite *ak* classique de plusieurs façons, les différences les plus importantes étant le recours à un micro-appariement pour la collecte des données de la période précédente au niveau de l'unité pour les panels communs et au calage par régression (comme dans le cas de l'estimateur gr) pour produire un ensemble de poids finals applicables à toutes les variables étudiées. Nous avons étudié trois versions de l'estimateur rc, c'est-à-dire l'estimateur principal sur l'estimateur rc2, c'est-à-dire l'estimateur avec prédicteur axé sur la variation (à cause du lissage résultant souhaité de la série d'estimation), nous avons constaté (bien que les résultats ne soient pas présentés ici) que les estimations du niveau de certaines

variables importantes peuvent encore être améliorées (comparativement aux estimations rc2) en incluant les prédicteurs axés sur le niveau correspondant. Donc, en pratique, une bonne stratégie consisterait à utiliser un mélange comportant principalement des prédicteurs axés sur la variation et quelques prédicteurs axés sur le niveau. La version de l'estimateur rc appliquée à l'heure actuelle pour l'EPAC, qui a été proposée par Fuller (1999), peut être formulée comme suit

$$C^{(rc)}_t = F_t + b^{(rc)}_t [(1 - \alpha)(\tilde{C}^{(rc)}_{t-1} - D^{*}_{t-1}) + \alpha(\tilde{C}^{(rc)}_{t-1} - D^{*}_{t-1})] \quad (6.1)$$

où α est fixé (à, disons, 1/3, mais, de façon générale, pourrait être particulier à y) et le coefficient $b^{(rc)}_t$ est calculé au moyen du système de régression généralisée comme s'il faisait partie de la classe rc d'estimations. On obtient une interprétation simple de l'équation (6.1) en la comparant à l'équation (3.7) de l'estimateur *ak*. Commençons par écrire (3.7) sous la forme

Tableau 3
Efficacité relative moyenne de l'estimateur rc par rapport à l'estimateur gr (Ontario, 1996)

Variable	Coeff.		Effic. (Niveau)		Effic. (Variation)	
	rc - unidimensionnel (niveau ou variation)	ak*	rc (uni- dimensionnel)	rc (multi- dimensionnel)	rc (uni- dimensionnel)	rc (multi- dimensionnel)
Personnes occupées	0,88 0,81 0,90	0,72 0,95	1,05	1,05	2,39	2,46
Chômeurs	0,60 0,53 0,65	0,50 0,69	1,12	1,12	1,31	1,33
Emploi Commerce	0,96 0,94 0,98	0,84 0,98	1,17	1,17	4,98	5,07
Emploi TRCO	0,95 0,93 0,97	0,87 0,98	1,37	1,37	7,47	7,52

Tableau 4
Efficacité relative moyenne des estimateurs ak* et rc par rapport à l'estimateur gr, Ontario, 1996 (ordonnaire c. différence)

Variable	ak* Coeff		Effic (ak*)		Effic (rc)		Variation	
	(ordonnaire)	(différence)	NA	0,91	2,55	2,32	0,97	4,88
Agriculture	(ordonnaire)	(différence)	NA	0,63	2,32	2,32	0,97	4,88
Agriulture	(ordonnaire)	(différence)	NA	0,63	2,32	2,32	0,97	4,88
NILF	(ordonnaire)	(différence)	NA	0,74	1,26	1,07	0,95	1,96
NILF	(différence)	(différence)	NA	1,21	1,07	1,07	1,95	2,01

Tableau 5
Relation entre les efficacités des estimations du niveau et de la variation pour rc (multidimensionnel)
par rapport à gr (Ontario, 1996)

Variable	Effic. Variation	Effic. Niveau	Effic. Variation/Effic. Niveau	(1 - p _{gr}) (1 - p _{rc})	p _{gr}	p _{rc}
Personnes occupées	2,46	1,05	2,34	2,65	0,77	0,91
Chômeurs	1,33	1,12	1,19	1,21	0,5	0,59
Emploi Commerce	5,07	1,22	4,16	3,80	0,79	0,95
Emploi TRCO	7,54	1,42	5,31	5,66	0,8	0,97

5.5 Séries chronologiques des estimations du niveau

Les figures 1(a) et (b) montrent les estimations du niveau de l'emploi pour l'Ontario pour la période allant de 1988 à 1996 obtenues au moyen de gr et rc avec et sans désaison-nalisation (par la méthode X11-ARIMA). Les figures 2(a) et (b) montrent l'emploi pour le groupe de branches d'acti-vité reprises sous « commerce ». Au niveau provincial, pour l'agrégation au niveau du groupe de branches d'activité, les séries (désaisonnalisées ou non) obtenues par régression généralisée (gr) et par régression composite (rc) sont semblables, parce qu'avant tout, les estimations par régression généralisée sont d'une haute précision. Par contre, au niveau du domaine « commerce », les séries sont assez différentes. (Notons que nous avons choisi cette série particulière parmi les nombreuses que nous avons examinées afin d'illustrer ici le scénario extrême pour ce qui est des écarts entre les séries gr et rc. Pour presque toutes les autres branches d'activité, les deux séries se recoupent assez fréquemment.) Puisque la série gr est

fortement instable, l'utilisation de l'estimateur rc peut donner lieu à un lissage considérable. Remarquons aussi qu'à cause du ratio signal-bruit élevé prévu, la série rc désaisonnalisée au niveau du domaine « commerce » paraît nettement plus lisse que la série gr; en fait, il y a fort peu d'écarts entre les séries gr désaisonnalisées et non désaison-nalisées. Nous constatons qu'on observe, en général, une série de périodes consécutives où l'estimation rc est soit plus grande soit plus faible que l'estimation gr. Cette situation est prévisible étant donné les valeurs élevées des coefficients $b^{(rc)}$ (tableau 3) et la forte corrélation sériale des deux séries (voir le tableau 5). Curensement, les points de renversement dans les séries gr et rc ont tendance à survenir (approximativement) au même moment, quoi qu'ils paraissent légèrement atténués dans le cas de rc à cause d'une corrélation sériale plus forte pour la série rc. Notons aussi que l'écart entre les deux séries serait plus faible si l'on incluait aussi les variables de contrôle pour les prédicteurs axés sur le niveau.

Tableau 1

Tableau 2

5.3 Estimation de la variation c. estimation du

niveau – efficacité de rc par rapport a gr

5.4 Estimation ponctuelle et écart-type de la

différence entre ce et gr

ak* des composantes d'un agrégat.

Le tableau 4 présente les pertes possibles d'efficacité pour les estimations obtenues par différentes méthodes de cohérence interne dans le cas de l'estimateur ak^* . Les résultats indiquent qu'en pratique, il faut se montrer prudent lors du choix des variables pour l'estimation par différence ou de l'utilisation de valeurs intermédiaires (de compromis) des coefficients pour l'estimation ak^* des composantes d'un agrégat.

variation pour estimer le niveau tout basser l'efficacité. Le tableau 3 donne une comparaison de l'estimateur ak^* (unidimensionnel et multidimensionnel) à l'estimateur ak^* . Les valeurs possibles des coefficients $b^{(1)(c2)}$ sur la période de 12 mois pour l'estimateur unidimensionnel $rc2$ sont résumées au moyen de la moyenne, du minimum et du maximum. On peut comparer ces valeurs aux coefficients optimaux correspondants pour ak^* . Les coefficients rc semblent représenter un compromis et se situent entre les valeurs du coefficient optimal pour le niveau et du coefficient optimal pour la variation. Les valeurs de l'efficacité de rc pour l'estimation de la variation sont ainsi dire équivalentes à celles obtenues pour ak^* , mais sont un peu plus faibles pour les estimations du niveau. Les gains d'efficacité sont faibles au niveau agréé pour lequel il existe des variables de contrôle pour l'estimateur gr mais sont grands pour les domaines sans contrôle pour gr .

le niveau pour estimer la variation et inversement. Les coefficients optimaux pour le niveau semblent donner, d'assez bons résultats pour les estimations de la variation, tandis que l'utilisation des coefficients optimaux pour la variation pour estimer le niveau font baisser l'efficacité.

Les résultats numériques se fondent sur les données de l'EPAC de 1996 pour l'Ontario (voir Singh, et coll. 1997). Les variables auxiliaires pour la régression généralisée sont les tableaux de population correspondants à 16 groupes d'âge-sexe, 11 régions économiques, 10 régions métropolitaines de recensement et 6 panels. Chaque total de contrôle de panel définit un sixième de la population de 15 ans et plus. Les nouveaux taux de contrôle (30 en tout) utilisés pour la régression composite correspondent uniquement aux prédicteurs axés sur la variation sont : les personnes occupées, les chômeurs et les personnes qui ne font pas partie de la population active selon le groupe d'âge (jeunes/vieux)-sexe (soit, en tout, 12 groupes). L'emploi selon le groupe de branches d'activité (soit, en tout, 16) et l'emploi selon le travail à temps plein ou à temps partiel (soit, en tout, 2). En fait, ces 30 nouvelles variables de contrôle se réduisent à 28 seulement, à cause de la

au groupe de renouvellement qui pourrait exister.

Le tableau 2 montre les coefficients optimaux (par exemple, k pour l'estimateur $a(2)$ et l'efficacité relative obtenus les coefficients optimaux par recherche par compar-

ativement en utilisant les données de 1996. (En pratique, cette recherche devrait se fonder sur des données antérieures).

Nous constatons que les gains d'efficacité peuvent être considérables lorsque l'on passe de ak à ak^* . Les coefficients optimaux diffèrent pour les estimations du niveau et de la variation. Les deux dernières colonnes sous les rubriques « niveau » et « variation » montrent la réduction de l'efficacité si l'on se sert des coefficients optimaux pour

la moyenne de l'estimateur composite sur ces 12

Pour les estimations du niveau, on calcule la corrélation entre l'estimation du niveau du mois courant (c 'est-à-dire F_t') et le prédicteur (c 'est-à-dire le prédicteur axé sur le niveau au niveau macro), tandis que pour l'estimation de la corrélation, on calcule la corrélation entre $C_{t-1}^{I-1} - D_{t-1}$ au niveau macro), et les prédicteurs $F'_t - C_{t-1}^{I-1}$. Comme il faut s'y attendre, la corrélation est négative, car l'estimation portant sur les paramètres communs est corrélée positivement à F_t' , mais exprimée avec un signe négatif dans le prédicteur. Rappelons que l'estimateur composite utilisé est ak pour les prédicteurs de niveau macro et ak^* pour les prédicteurs de niveau micro.

Le tableau I montre que, pour les quatre variables importantes (personnes occupées, chômeurs, personnes occupées dans le secteur du commerce et des communications (TRCO)), pour chaque prédicteur axé sur le niveau ou axé sur la variation, le prédicteur de niveau donne lieu à une plus forte corrélation que le prédicteur de niveau

et la colonne correspondante de la matrice $X(t)^*$ comprend les $n(t)$ valeurs de $y_k^g(t) + (1 - \gamma)^{-1} (y_k^g(t) - 1) I^{k \in s(t)-1}$. Lorsque la matrice $X(t)^*$ est définie, on peut se servir du système de régression généralisée pour calculer les poids $w_{rc}(t, k)$ par calage comme dans (3.2). On notera que l'on peut utiliser les poids calés $w_{rc}(t, k)$ pour estimer toutes les variables étudiées, mais qu'ils ne dépendent explicitement que de l'ensemble clé de variables étudiées choisies comme nouveaux prédicteurs d'après les données antérieures corrélées. Notons aussi que, même si on a défini l'estimateur rc donné par l'équation (3.6) comme étant l'estimateur gr corrigé par régression pour tenir compte des nouveaux prédicteurs, pour faciliter le calcul, il est commode de procéder à un calage par régression généralisée sur les poids de sondage lorsque l'on considère simultanément les anciens et les nouveaux contrôles de calage. De cette façon, le calcul de l'estimateur rc multivarié ne diffère peu de celui de l'estimateur rc univarié. Autrement, on pourrait calculer l'estimateur rc sous forme d'estimateur corrigé par régression généralisée comme dans (3.6), mais les coefficients de régression partielle prédicteurs seraient des coefficients de régression partielle et, par conséquent, n'auraient pas la forme type des coefficients de régression généralisée.

Enfin, nous constatons que, dans le cas de l'estimation composite, on s'attendrait à observer des gains d'efficacité plus importants pour les estimations de la variation $(C_i - C_{i-1}, F_i - F_{i-1})$ que pour les estimations du niveau $(C_i, C_{i-1}, F_i, F_{i-1})$. Pour montrer ceci, considérons une identité simple: $V(F_i - F_{i-1}) = V(F_i) + V(F_{i-1}) - 2\text{Cov}(F_i, F_{i-1})$. Typiquement, $V(F_i) \approx V(F_{i-1}) = \sigma_{gr}^2$ (disons), alors l'équation qui précède peut être réduite à $V(F_i - F_{i-1}) \approx 2\sigma_{gr}^2(1 - \rho_{gr})$. Pareillement, $V(C_i - C_{i-1}) \approx 2\sigma_{rc}^2(1 - \rho_{rc})$. Donc, l'efficacité de l'estimation de la variation est approximativement égale à l'efficacité de l'estimation du niveau multipliée par $(1 - \rho_{gr}) / (1 - \rho_{rc})$. Il s'ensuit que si les nouveaux prédicteurs pour l'estimation composite augmentent considérablement la corrélation (positive) entre C_i et C_{i-1} , alors l'efficacité de l'estimation de la variation sera nettement supérieure à celle de l'estimation du niveau.

4. ESTIMATION DE LA VARIANCE

À l'heure actuelle, dans le cas de l'EPAC, on applique la méthode du jackknife avec élimination d'UPB pour calculer la variance de l'estimation par régression généralisée. La méthode du jackknife est valide (pour les enquêtes transversales) si les estimations au niveau de l'UPB ont les mêmes moyennes et variances et que l'on peut traiter la sélection des UPB comme un tirage avec remise. Si la sélection des UPB se fait sans remise, l'estimation de la variance devient prudente si, comme cela est généralement le cas, la covariance (commune) entre les estimations au

composites ak^* sont les mêmes que les poids composites ak de Fuller (1990) où les estimateurs composites pour un ensemble de variables y importantes servent de contrôles supplémentaires lors de la régression généralisée ordinaire pour obtenir un ensemble de poids calés finals. Ceci permet de calculer l'estimateur composite ak comme s'il agissait d'un estimateur par calage.) Les différences principales entre les divers estimateurs décrits plus haut tiennent à la définition des coefficients de régression (optimaux c , sous-optimaux) et à celle des prédicteurs (utilisation de données antérieures de niveau micro c , macro). On peut obtenir les cas spéciaux des estimateurs composites susmentionnés décrits dans Singh, et coll. (1997) en n'utilisant que l'un des deux prédicteurs. Pour $C_i^{(ak)}$, si $a = 0$ (c'est-à-dire si l'on se sert uniquement d'un prédicteur k bien connu que l'on peut qualifier d'estimateur $ak2$ dans le présent contexte. Si $a = k$, c'est-à-dire si l'on utilise uniquement un prédicteur axé sur le niveau, nous obtenons un nouvel estimateur composite $C_i^{(ak1)}$ que l'on peut appeler estimateur $ak1$. De la même façon, pour $C_i^{(ak*)}$, nous obtenons deux nouveaux estimateurs composites supplémentaires, ak^*1 et ak^*2 . Pour $C_i^{(rc)}$, si nous utilisons uniquement un prédicteur axé sur le niveau, nous obtenons l'estimateur $rc1$, appelé précédemment MR1 dans Singh et Merikouris (1995). Si l'on se sert uniquement de prédicteurs axés sur la variation, nous obtenons l'estimateur $rc2$ appelé antérieurement MR2 dans Singh, et coll. (1997).

Comme nous l'avons mentionné plus haut, l'estimateur rc est calculé avec un estimateur gr de (3.1) et, par conséquent, peut être exprimé sous la forme $\hat{y}_{y(t)}(t) = \sum_{k \in s(t)} y_k(t) w_{rc}(t, k)$. Nous étendons la matrice $X(t)$ à la matrice $n(t) \times (p + 2q)$, soit $X(t)^*$, où $2q$ représente le nombre de nouveaux prédicteurs, le facteur 2 signifiant que l'on se sert de deux prédicteurs, l'un axé sur le niveau et l'autre sur la variation. Les tableaux de contrôle (aléatoires) $C_{i-1}^{(rc)}$ qui correspondent à l'ensemble clé de variables y de la période $i - 1$ sélectionnés pour l'estimation composite sont traités comme des constantes (durant le calcul du coefficient de régression) au même titre que les autres contrôles gr (non aléatoires). Maintenant, puisque le prédicteur axé sur le niveau peut être écrit sous la forme

$$D_{i-1}^* = (1 - \gamma)^{-1} \sum_{k \in s(t)-1} y_k(t-1) w_{gr}(t, k) = (1 - \gamma)^{-1} \sum_{k \in s(t)} y_k(t-1) I^{k \in s(t)-1} w_{gr}(t, k) \quad (3.8)$$

la colonne de la matrice $X(t)^*$ qui correspond à ce prédicteur comprend $n(t)$ valeurs de $(1 - \gamma)^{-1} y_k(t-1) I^{k \in s(t)-1}$. De la même façon, le prédicteur axé sur la variation peut s'écrire sous la forme

$$F_i^* + D_{i-1}^* - B_i^* = \sum_{k \in s(t)} y_k(t) + (1 - \gamma)^{-1} (y_k(t-1) - y_k(t-1) I^{k \in s(t)-1}) w_{gr}(t, k) \quad (3.9)$$

de C_t est réduite au minimum. Comme nous l'avons mentionné plus haut, habituellement a est plus petit que k . Pour définir les deux nouvelles fonctions prédictives nulles susmentionnées en se fondant sur les données antérieures, on commence par former deux estimateurs de $\tau_t(t-1)$: l'un est \bar{D}_{t-1} , fondé sur les panels de la période $t-1$ conservés (c'est-à-dire le sous-échantillon de la période $t-1$ apparié à l'échantillon de la période t) et l'autre est $F_t + (D_{t-1} - B_t)$ qui est l'estimateur gr à la période t , corrigé pour la variation de $t-1$ à t estimée d'après l'échantillon commun. De toute évidence, si le plan de sondage ne prévoit aucun chevauchement de panel, les fonctions prédictives nulles perdent leur signification et l'estimation composite ne modifie pas F_t . Pareillement, si le chevauchement est complet, $B_t = F_t$ et, de nouveau, l'estimation composite n'a aucun effet sur F_t . À première vue, cela paraît contraire à l'intuition, puisque les données de la période précédente (y_{t-1}) sont corrélées à celles de la période courante (y_t) à cause du chevauchement de l'échantillon. Cependant, le chevauchement complet se résume, en principe, à recueillir sur y_t auprès d'un échantillon unique, des données multidimensionnelles dont les éléments correspondent à y à diverses périodes. Selon cette analogie, il n'y a aucune possibilité d'amélioration (dans le cadre de référence axée sur le plan de sondage), puisqu'il n'existe aucun échantillon plus grand pouvant fournir des données supplémentaires. Dans le cas où il n'y a aucun chevauchement, l'information supplémentaire existe, mais elle n'est pas utile, car elle n'est pas corrélée. On notera, cependant, qu'au premier degré, les UPB (unités primaires d'échantillonnage) de l'EPAC restent communes pendant plusieurs années avant d'être supprimées de l'échantillon. Par conséquent, les gains d'efficacité dus au chevauchement partiel sont réalisés principalement grâce à la réduction de la composante de la variance liée à l'échantillonnage de deuxième degré.

De surcroît, soulignons que l'estimateur $C_t^{(ak)}$ se fonde sur des données antérieures de façon unidimensionnelle, puisque, pour la variable étudiée y , on se sert uniquement de l'information y_{t-1} provenant de la période précédente. Si l'on se sert pour la variable étudiée y de prédicteurs fondés sur plusieurs variables des périodes précédentes, comme $y_{t-1}^1, y_{t-1}^2, \dots, y_{t-1}^k$, alors l'estimation composite devient multidimensionnelle. Cependant, le choix optimal des coefficients (a, k) pour le cas multidimensionnel peut être assez fastidieux.

La catégorie d'estimateurs par régression composite (rc) est donnée par

$$C_t^{(rc)} = F_t + b_{t^{(rc)}} \left(\bar{C}_{t-1}^{(rc)} - \bar{D}_{t-1}^* + B_{t-1} - F_t \right) + a_{t^{(rc)}} \left(\bar{C}_{t-1}^{(rc)} - \bar{D}_{t-1}^* \right) \quad (3.6)$$

où $\bar{C}_{t-1}^{(rc)}$ représente l'estimateur $t-1$ pour la variable étudiée y après que les poids calés à la période $t-1$ soient

La variable de contrôle $\bar{C}_{t-1}^{(ak)}$ représente l'estimateur par calage à la période ($t-1$) pour y après que les poids composites ak^* soient de nouveau calés sur les variables de contrôle utilisées pour la stratification *a posteriori* par régression généralisée à la période t . Ici les poids

$$C_t^{(ak)} = F_t + (k^* - a^*) \left(\bar{C}_{t-1}^{(ak)} - \bar{D}_{t-1}^* + B_{t-1} - F_t \right) + a^* \left(\bar{C}_{t-1}^{(ak)} - \bar{D}_{t-1}^* \right) \quad (3.7)$$

de nouveau calés sur les contrôles utilisés pour la stratification *a posteriori* dans le cas de la régression généralisée (gr) à la période t . Donc, $\bar{C}_{t-1}^{(rc)}$ est une estimation du total de la population à la période t pour la variable y à la période $t-1$. Dans \bar{D}_{t-1}^* , l'astérisque signifie que ce terme est fondé sur le sous-échantillon de la période t apparié à l'échantillon de la période $t-1$, mais que l'on a utilisé les poids gr à la période t , puisque les valeurs de y provenant de la période $t-1$ sont augmentées par micro-appariement de façon à ce qu'elles correspondent aux valeurs de l'échantillon à la période t . (À noter que \bar{D}_{t-1}^* contient, en général, des valeurs imputées et qu'il peut être entaché d'un biais dû à l'imputation. Pour le diagnostic et le rajustement de ce biais, voir la section 6.) Les coefficients $b_{t^{(rc)}}$ et $a_{t^{(rc)}}$ sont calculés de la même façon que pour l'estimateur gr donné par l'équation (3.1). Ces des précisions supplémentaires sont données plus loin. Cependant, comme (a, k), ils sont particuliers à y , et, dans le cas de l'estimation multidimensionnelle, dépendent de l'ensemble clé de variables étudiées provenant de la période précédente pour servir de nouveaux contrôles, mais ils peuvent être calculés facilement puisqu'ils sont sous-optimaux. Donc, dans le cas de l'estimation rc, il est assez facile d'introduire un plus grand nombre de prédicteurs. Les prédicteurs ($C_{t-1} - \bar{D}_{t-1}$) et ($C_{t-1} - \bar{D}_{t-1} + B_{t-1} - F_t$) peuvent être qualifiés, respectivement, de prédicteur axé sur le niveau et de prédicteur axé sur la variation comme dans Singh, Kennedy, Wu et Brisebois (1997). Il en est ainsi non seulement parce que le premier est la différence de deux estimations de niveau et le second, la différence de deux estimations de la variation, ($C_{t-1} - F_t$) et ($\bar{D}_{t-1} - \bar{B}_{t-1}$), mais aussi parce que le premier a tendance à augmenter fortement l'efficacité de l'estimation du niveau comparativement au gain que l'on peut obtenir en présence du second et, pareillement, que le second augmente considérablement l'efficacité de l'estimation de la variation comparativement à celle que l'on peut obtenir en présence du premier.

Nous pouvons appliquer l'utilisation des microdonnées de la période précédente pour obtenir les nouveaux prédicteurs pour l'estimation rc à l'estimateur ak^* donné par un nouvel estimateur ak^* donné par

3. ESTIMATEURS COMPOSITES : NOUVEAUX ET ANCIENS

Nous commençons par l'estimateur transversal du total $\tau_y(t)$ à la période t défini comme étant l'estimateur par régression généralisée gr, qui est donné par

$$\hat{\tau}_y^{gr}(t) = \sum_{k \in s(t)} y_k^t(t) w_{gr}^k(t, k), \quad (3.1)$$

$$= d(t, k) [1 + x_k^t(t)' \Delta(t) X(t)]^{-1} (\tau_x^t(t) - \hat{\tau}_x^t(t)), \quad (3.2)$$

où les $d(t, k)$ sont les poids de sondage initiaux corrigés pour la non-réponse, $x_k^t(t)$ est un vecteur p de covariables utilisées pour le calage (ou la stratification a posteriori), $X(t)$ est la matrice $n(t) \times p$ des observations de $x, n(t)$ est la taille de l'échantillon, $\Delta(t)$ est $\text{diag}(d(t, k))$, $\tau_x^t(t)$ est le vecteur p connu des contrôles utilisés pour le calage et $\hat{\tau}_x^t(t)$ est le vecteur correspondant des estimations avec facteur d'extension fondées sur les poids d . Si l'on reprend la notation F^t, B^t , et B^t de la section précédente, F^t peut être considéré comme l'estimateur gr (3.1) et B^t , comme l'estimateur gr fondé sur les panels déjà existants donnés par

$$\bar{B}^t = (1 - \gamma)^{-1} \sum_{k \in s(t|t-1)} y_k^t(t) w_{gr}^k(t, k), \quad (3.3)$$

où $s(t|t-1)$ est le sous-échantillon de la période t apparié à l'échantillon de la période $t-1$. L'estimateur B^t est également un estimateur gr qui peut être représenté par

$$C^{t(ak)} = F^t + k(C^{t-1(ak)} - \bar{D}^{t-1} + B^t - F^t) + a(F^t - \bar{B}^t) = F^t + (k-a)(C^{t-1(ak)} - \bar{D}^{t-1} + B^t - F^t) + a(C^{t-1(ak)} - \bar{D}^{t-1}). \quad (3.5)$$

Ici, on obtient les coefficients a, k pour l'estimation du niveau par régression optimale de F^t sur les deux fonctions prédictives nulles, basées sur les données antérieures, à savoir $C^{t-1(ak)} - (F^t + \bar{D}^{t-1} - B^t)$ et $C^{t-1(ak)} - \bar{D}^{t-1}$. Donc, a et k dépendent du plan d'échantillonnage, ainsi que de la variable étudiée y ; plus précisément, ils ne sont même pas identiques pour les estimations du niveau et de la variation pour une même variable y . Pour l'estimation de la variation, c'est la régression de $F^t - C^{t-1(ak)}$ et non de F^t que l'on calcule de façon optimale sur les prédicteurs. En pratique, on estime a, k en exploitant par compartiment l'intervalle (0,1) afin de trouver les valeurs pour lesquelles la variance

$$C^t = (1 - b^t - a^t) F^t + b^t (C^{t-1} + \bar{B}^t - \bar{D}^{t-1}) + a^t (F^t + C^{t-1} - \bar{D}^{t-1}). \quad (2.8)$$

La justification heuristique qui précède s'appuie sur des considérations de réduction de la variance si l'on suppose qu'il n'y a aucun biais dû au groupe de renouvellement lorsque l'on combine trois estimateurs non biaisés du total de population à la période t . Par contre, l'existence d'un biais dû au groupe de renouvellement introduit dans les trois estimateurs un biais dont la grandeur et le signe peuvent différer selon l'estimateur et, dans ce cas, l'estimation composite rajuste chacun des estimateurs de sorte que la valeur corrigée pour chacun soit égale à une valeur commune donnée par l'estimateur composite (par exemple, dans le cas de deux estimateurs θ_1 et θ_2 de θ , la combinaison linéaire $\lambda_1 \theta_1 + (1 - \lambda_1) \theta_2$ peut s'écrire sous la forme $\theta_2 + \lambda (\theta_1 - \theta_2) + (1 - \lambda) \theta_2$, ce qui sous-entend que les deux estimateurs originaux sont corrigés de façon appropriée pour qu'ils convergent vers une valeur commune). Le poids relatif appliqué lorsque l'on combine les trois estimateurs dépend du critère de variance minimale. Idéalement, il devrait se fonder sur le critère de l'EQM minimale, mais il est difficile de maîtriser le biais, car on ne peut l'estimer. L'estimateur composite n'est manifestement pas dépourvu de biais et l'on peut simplement spéculer que le biais global de l'estimateur est réduit par le recours à un estimateur composite. Selon le même raisonnement si, inversement, on utilise une régression sous-optimale pour créer l'estimateur composite (comme dans l'estimation rc, voir la section suivante), l'effet de l'estimation composite est de rajuster les poids d'échantillonnage dans l'échantillon complet (qui sont généralement des poids gr) de sorte que $F^t - (B^t - \bar{D}^{t-1})$ et \bar{D}^{t-1} avec les poids rajustés, devient égal à C^{t-1} ; les valeurs de C^{t-1} servent de nouvelles valeurs de contrôle à l'étape du calage. Il s'agit là d'un autre moyen de rajuster les trois estimateurs sur une valeur commune, mais, de nouveau, le biais introduit par l'estimateur composite est inconnu. L'exposé de deux perspectives sous lesquelles on peut examiner l'estimation composite présente des points communs avec l'effet double de la stratification a posteriori qui réduit à la fois la variance et le biais (de couverture); à ce sujet, consulter Singh et Folsom (2000).

mois courant; voir l'équation (2.8) à la fin de la présente section pour l'expression réelle.) Maintenant, pour montrer le lien avec l'estimation composite bien connue définie à la section suivante, posons $0 < a_i, b_i < 1$, de sorte que $\delta_{1i} = 1 - b_i, \delta_{2i} = 1 - b_i - a_i$. Nous avons

$$C_i' = C_{i-1} + (B_i' - \bar{D}_{i-1}) \gamma (B_i' - \bar{B}_i') \quad (2.6)$$

$$+ (1 - b_i - a_i) (\bar{D}_{i-1} - C_{i-1}).$$

Il est intéressant de noter que si $b_i' = 0$, il n'y aura aucun amenuisement de l'effet dû au nouveau panel et la série C se rapprochera, en principe, de la série F , c'est-à-dire que le lissage sera moins important et que les deux séries se recouperont plus souvent. Si $a_i' = 0$, l'effet dû au panel de la période précédente qui a été supprimé, représente par $(\bar{D}_{i-1} - C_{i-1})$, est moins amorti. Autrement dit, la série F sera lissée plus fortement et, en principe, les deux séries se recouperont moins fréquemment. Enfin, si $a_i', b_i' > 0$, le comportement de la série C par rapport à la série F sera compris entre les deux comportements susmentionnés. Qui plus est, si la valeur de b_i' est grande (proche de 1), le lissage de la série F sera assez prononcé, car les effets du nouveau panel et du panel supprimé seront tous deux fortement amortis. Dans ces situations, on s'attendrait à observer des écarts prolongés entre les séries F et C au cours du temps avant qu'elles ne se recourent. On notera que les parties du terme $\gamma (B_i' - \bar{B}_i')$ qui sont amorties sur les périodes $t, t + 1, \dots$ diminuent à mesure que t augmente. Elles sont données par $b_i' \gamma (B_i' - \bar{B}_i'), (b_{i+1}' + a_{i+1}') b_i' \gamma (B_i' - \bar{B}_i'), \dots$. Pareillement, les parties amorties de $(\bar{D}_{i-1} - C_{i-1})$ sont

$$(b_i' + a_i') (\bar{D}_{i-1} - C_{i-1}), (b_{i+1}' + a_{i+1}') (b_i' + a_i') (\bar{D}_{i-1} - C_{i-1}), \dots.$$

De toute évidence, si b_i' est grand, l'amortissement complet prend plusieurs périodes. Cependant, comme on l'a expliqué plus haut, cette situation n'introduira pas de biais, car les effets amortis sont des fonctions nulles si l'on suppose qu'il n'y a pas de biais dû au groupe de renou-

vèlement.

L'expression (2.6) peut être écrite sous une forme plus connue de l'estimateur composite, comme suit :

$$C_i' = C_{i-1} + (B_i' - \bar{D}_{i-1}) + (1 - b_i') (F_i' - \bar{B}_i') \quad (2.7a)$$

$$+ (1 - b_i') (\bar{D}_{i-1} - C_{i-1}) + a_i' (C_{i-1} - \bar{D}_{i-1}) \quad (2.7b)$$

$$= F_i' + b_i' [C_{i-1} - (F_i' - \bar{B}_i') + a_i' (C_{i-1} - \bar{D}_{i-1})] \quad (2.7c)$$

$$= F_i' + (b_i' + a_i') (C_{i-1} - \bar{D}_{i-1}) + \bar{B}_i' - F_i' \quad (2.7d)$$

L'expression (2.7d) correspond à l'estimateur ak (voir la section suivante) si $a_i' = a$ et $b_i' + a_i' = k$. En pratique, on

l'estimateur C_i' demeure non biaisé. L'espérance de la différence $(F_{i-1}' - C_{i-1}')$ est nulle si l'on suppose que C_{i-1}' et F_{i-1}' ne sont pas biaisés (ce qui est le cas si l'on suppose qu'il n'existe aucun biais dû au groupe de renouvellement) et, par conséquent, leur amortissement partiel n'aura pas d'influence sur l'absence de biais des futures estimations C_i' . Cependant, en réalité, l'espérance de la différence $F_{i-1}' - F_{i-1}$ n'est pas nulle et il convient d'être prudent lors de l'amortissement d'une de ses parties. Notons que

$$F_i' - F_{i-1}' = (B_i' - \bar{D}_{i-1}) + \gamma (B_i' - \bar{B}_i') + \gamma (\bar{D}_{i-1} - D_{i-1}). \quad (2.3)$$

Le premier terme du deuxième membre de l'équation est l'estimation de la variation fondée sur les panels communs, tandis que les deuxième et troisième représentent l'effet du panel ajouté et du panel retiré aux périodes t et $t - 1$, respectivement. Les deux derniers termes sont des fonctions nulles (c'est-à-dire que leur espérance est nulle), mais le premier ne l'est pas (heureusement, ce premier terme devrait être stable, car il s'agit de la différence entre deux estimations fortement corrélées). Par conséquent, ce sont les deuxième et troisième termes qui doivent être amortis. Maintenant, écrivons l'équation (2.2) sous la forme

$$F_i' = C_{i-1} + (B_i' - \bar{D}_{i-1}) + \gamma (B_i' - \bar{B}_i') \quad (2.4)$$

$$+ [\gamma (\bar{D}_{i-1} - D_{i-1}) + (F_{i-1}' - C_{i-1})]$$

$$+ [(D_{i-1}' - F_{i-1}') + (F_{i-1}' - C_{i-1})]$$

$$= C_{i-1} + (B_i' - \bar{D}_{i-1}) + \gamma (B_i' - \bar{B}_i') + (\bar{D}_{i-1} - C_{i-1}).$$

et définissons deux facteurs d'amortissement δ_{1i}, δ_{2i} , dont la valeur est comprise entre 0 et 1, puis définissons la série lissée $\{C_i'\}$ comme étant

$$C_i' = C_{i-1} + (B_i' - \bar{D}_{i-1}) + \delta_{1i} \gamma (B_i' - \bar{B}_i') + \delta_{2i} (\bar{D}_{i-1} - C_{i-1}). \quad (2.5)$$

Le terme qui contient δ_{1i} dans l'équation (2.5) représente la réduction de l'effet du nouveau panel à la période t dont C_i' essaye de rendre compte, tandis que le terme qui contient δ_{2i} se rapporte approximativement à la réduction de l'effet du panel de la période précédente $(t - 1)$ que C_i' essaye de compenser à la période t . En outre, il serait souhaitable de poser $\delta_{2i} < \delta_{1i}$ pour que la série $\{C_i'\}$ suive mieux la série $\{F_i'\}$, de sorte qu'elles soient caractérisées toutes deux par une même tendance au cours du temps, ce qui revient à accorder plus d'importance à l'effet du nouveau panel à la période courante qu'à celui du panel de la période précédente qui a été supprimé. (En fait, une justification rigoureuse, dans des conditions assez générales, de la contrainte $\delta_{2i} < \delta_{1i}$ provient de considérations d'optimalité en vertu desquelles la variance de C_i' est réduite au minimum pour obtenir la meilleure combinaison linéaire possible des trois estimateurs non biaisés, $F_i', C_{i-1}' + B_i' - \bar{D}_{i-1}$ et $F_i' + C_{i-1}' - \bar{D}_{i-1}$ du total de population du

au nombre moyen de ménages qui déménagent hors de ces logements à la période t , alors, même si les personnes qui ont déménagé ont d'autres caractéristiques d'emploi que celles qui n'ont pas déménagé, l'imputation de données pour les personnes qui ont déménagé ne devrait, en principe, introduire aucun nouveau biais, puisque l'on tient compte de la situation d'emploi durant le mois courant pour les autres covariables.

Dans nos conclusions, à la section 6, nous proposons une méthode pour diagnostiquer l'effet de cette imputation. Cet effet peut être important dans le cas d'enquêtes où la fraction de valeurs qui manquaient à la période précédente est importante pour les répondants du mois courant qui font partie de l'échantillon chevauchant. Un moyen simple de résoudre ce problème consisterait à remanier la question-naire de sorte que l'application informatique d'ITAO couramment de nos jours) indique à l'intervieur, lorsqu'il procède à l'interview le deuxième mois ou les mois suivants, si le répondant avait répondu ou non le mois précédent. Dans la négative, l'intervieur poserait alors quelques questions supplémentaires afin de déterminer la situation d'emploi du répondant le mois précédent. Cette solution est comparable à la méthode proposée par Hansen-Hurwitz-Madow pour des enquêtes répétées avec échantillons entièrement non chevauchants, où des questions sont posées à chaque répondant concernant la période courante et la période précédente (voir Cochran 1977, page 355).

Le plan de l'article est le suivant. À la section 2, nous présentons une justification heuristique de l'utilisation du concept d'amortissement pour expliquer pourquoi on produit le lissage souhaité de la série d'estimations. À la section 3, nous définissons divers estimateurs et décrivons leur calcul au moyen du système de régression généralisée gr. Nous proposons aussi une nouvelle version de l'estimateur ak , représentée par ak^* . Cet estimateur comprend des prédicteurs provenant du mois précédent obtenus par micro-appariement et devrait, en principe, permettre de réaliser des gains d'efficacité importants. À la section 4, nous considérons l'estimation de la variance selon la méthode du jackknife utilisée à l'heure actuelle. À la section 5, nous présentons une comparaison empirique des estimateurs fondée sur des données de l'EPAC de 1996 pour l'Ontario. Enfin, nous présentons nos conclusions à la section 6.

2. LISSAGE DES SÉRIES PAR ESTIMATION COMPOSITE : JUSTIFICATION HEURISTIQUE

Nous présentons ici une justification heuristique intéressante (fondée sur le concept d'amortissement plutôt que sur le concept de réduction) du fait que l'on s'attend à ce que l'estimation composite produise un lissage de la

série d'estimations. (Si l'on s'appuie uniquement sur le concept de réduction, la série sera lissée, mais ne recouvrera pas assez souvent la série originale. Par contre, en appliquant le concept d'amortissement, on rend compte graduellement au cours du temps de la partie restante après la réduction, ce qui permet le recoupement plus fréquent de la série lissée et de la série originale.) Considérons un plan de sondage avec renouvellement de panel comparable à celui de l'EPAC et représentons par γ la fraction des panels retirés de l'échantillon; dans le cas de l'EPAC, γ est égal à 1/6. Représentons l'estimateur transversal (classiquement, gr) à la période t fondé sur tous les panels (c'est-à-dire l'échantillon complet) par F_t , l'estimateur fondé sur le nouveau panel uniquement (c'est-à-dire le panel ajouté à l'échantillon) par B_t , et celui fondé sur les panels déjà existants (c'est-à-dire le sous-échantillon qui, à la période t chevauche l'échantillon de la période précédente $t-1$) par D_t . Pareillement, représentons l'estimateur fondé uniquement sur le panel supprimé (c'est-à-dire retiré de l'échantillon) par D_t , et celui fondé sur les panels non supprimés (c'est-à-dire le sous-échantillon qui, à la période $t-1$, chevauche l'échantillon de la période courante t) par D_{t-1} . Nous obtenons

$$F_t = \gamma B_t + (1 - \gamma) D_t \quad (2.1a)$$

$$F_{t-1} = \gamma D_{t-1} + (1 - \gamma) D_{t-1} \quad (2.1b)$$

Supposons que la série $\{F_t\}$ est trop instable et que nous voulons la lisser. Dans la suite, nous supposons qu'aucun biais n'est introduit par le groupe de renouvellement (Ballar 1975), c'est-à-dire que la valeur prévue est la même pour divers groupes de renouvellement. Donc, F_t est non biaisé, mais peut être instable. Il s'agit-là des conditions classiques de l'estimation composite pour laquelle diverses estimations non biaisées sont combinées de façon optimale pour obtenir une estimation plus efficace. Pour une autre perspective sur l'estimation composite en cas de biais dû au groupe de renouvellement, consulter la discussion à la fin de la présente section. Maintenant, représentons la série lissée par $\{C_t\}$, et considérons l'identité :

$$F_t = C_{t-1} + (F_t - F_{t-1}) + (F_{t-1} - C_{t-1}) \quad (2.2)$$

Nous pouvons interpréter la relation qui précède comme suit. L'estimation C_{t-1} à la période $t-1$ est corrigée en tenant compte de la fluctuation $(F_t - F_{t-1})$ à la période t dans la série F_t et de l'écart existant $(F_{t-1} - C_{t-1})$ à la période $t-1$. Si nous définissons C_t après les ajustements complets pour ces deux différences, C_t sera identique à F_t et il n'y aura aucun lissage de la série F_t . Ces résultats laissent entendre qu'on ne devrait expliquer que partiellement les ajustements pour les différences $(F_t - F_{t-1})$ et $(F_{t-1} - C_{t-1})$ lorsque la série C passe de C_{t-1} à C_t . Les parties restantes des différences devraient être amorties graduellement sur les périodes futures. Tous ces ajustements devraient être effectués en s'assurant que

Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel

AVINASH C. SINGH, BRIAN KENNEDY et SHIYING WU¹

RÉSUMÉ

Nous considérons l'estimation composite par régression introduite par Singh (1994, 1996), qui a été appelée au début « estimation composite par régression modifiée » et dont une version (proposée par Fuller (1999) est appliquée à l'Enquête sur la population active du Canada (EPAC) depuis janvier 2000. L'estimateur par régression composite (rc) comporte plusieurs améliorations comparativement à l'estimateur par régression généralisée (gr) et l'estimateur composite *ak* bien connu de Gurney-Daly. Les caractéristiques principales de l'estimateur rc sont les suivantes : a) il augmente considérablement l'efficacité de l'estimation du niveau et de la variation pour les variables étudiées importantes, ce qui produit des séries d'estimations moins instables; b) comme l'estimateur gr, il est calculé de la même façon qu'un estimateur par calage, de sorte que soient satisfaites les valeurs de contrôle utilisées habituellement pour la stratification *a posteriori* dans gr, ainsi que les nouvelles valeurs de contrôle qui correspondent aux variables corrélées provenant de la période d'observation précédente; c) il maintient la cohérence interne des estimateurs sans qu'il soit nécessaire de calculer les estimations partielles par différence. Les innovations les plus importantes qui caractérisent la classe des estimateurs rc consiste à : a) utiliser des matrices de covariances de travail pour estimer les fonctions au lieu de recourir à la modélisation d'une superpopulation pour définir les coefficients de régression des prédicteurs inclus dans l'estimateur gr, b) traiter les contrôles aléatoires (ceux fondés sur les variables corrélées importantes des périodes antérieures) comme étant des constantes, tout en calculant les coefficients de régression de la même façon que pour l'estimation à deux phases et comme étant justifiés par le concept de la matrice de covariances de travail et c) se fonder sur le micro-appariement pour obtenir des données auxiliaires de même niveau que les estimations antérieures pour obtenir une plus forte corrélation avec les variables étudiées à la période courante. Secondairement, nous recommandons d'utiliser une nouvelle version de l'estimateur *ak* qui s'appuie sur le macro-appariement basé sur les prédicteurs des périodes antérieures plutôt que sur les microdonnées classiques afin d'augmenter les gains d'efficacité. L'article présente aussi une justification heuristique intéressante de la caractéristique de lissage des estimations composites fondées sur le concept de l'amortissement. Enfin, nous présentons les résultats empiriques de la comparaison de divers estimateurs basés sur les données de l'EPAC de 1996 pour l'Ontario.

MOTS CLÉS : Régression généralisée; régression modifiée; fonctions d'estimation; calage par régression.

1. INTRODUCTION

Dans le cas des enquêtes répétées avec chevauchement partiel des échantillons, il est bien connu (voir, par exemple Cochran 1977, chapitre 12) que l'on peut améliorer les estimations ponctuelles du niveau pour une période et de la variation entre deux périodes par régression de l'estimateur transversal ordinaire (habituellement, estimé de la régression ou simplement estimateur d'Horvitz-Thompson) sur les nouveaux prédicteurs fournis par les observations corrélées faites sur le sous-échantillon chevauchant de la période précédente. Ces méthodes d'estimation rentrent dans la catégorie de l'estimation composite dont une version simple, appelée estimateur composite *k*, a été proposée il y a quelque temps par Hansen, Hurwitz et Madow (1953) et étudiée plus en détail par Rao et Graham (1964), Binder et Hidiroglou (1988), quant à eux, font une excellente revue des articles qui traitent de l'estimation dans le cas d'enquêtes répétées. Il convient de souligner que l'aggrégation des estimations sur plusieurs périodes cause une perte d'efficacité, à cause de la plus forte corrélation positive entre les estimations composites ponctuelles

successives. Néanmoins, il s'agit probablement d'un faible prix à payer, car ce n'est pas la précision de l'agrégat, mais celle des estimations du niveau et de la variation qu'il faut améliorer. L'estimateur composite *ak* de Gurney et Daly (1965) est une version améliorée de l'estimateur composite *k* obtenue grâce à une réduction supplémentaire de la variance et pour laquelle une autre justification plus simple a été fournie par Wolter (1979). L'estimateur composite considéré ici a été développé dans le contexte de l'Enquête sur la population active du Canada (EPAC). Cette dernière est une enquête mensuelle fondée sur un plan de sondage avec renouvellement de panel et comptant six panels. Pendant deux mois consécutifs, cinq sixième des ménages sélectionnés fortement (proposée par Fuller 1999) des estimateurs composites introduits par Singh (1994, 1996) et appelés au départ estimateurs « composites par régression modifiée », que nous dénommerons ici simplement estimateurs « composites par régression » ou estimateurs rc. Avant janvier 2000, on utilisait pour l'EPAC les estimateurs par régression

¹ Avinash C. Singh et Shiyong Wu, Statistical Research Division, Research Triangle Institute, Research Triangle Park, N.C. 27709-2194, U.S.A.; Brian Kennedy, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

- TUCKER, C., CASADY, R. et LEPKOWSKI, J. (1992). Sample allocation for stratified telephone sample designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 291-296.
- TURNER, C.F., FORSYTH, B.H., O'REILLY, J.M., COOLEY, P. C., SMITH, T.K., ROGERS, S.M. et MILLER, H.G. (1998). Automated self-interviewing and the survey measurement of sensitive behaviors. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et coll. - eds.). New York: John Wiley and Sons, 455-473.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- WAKSBERG, J. (1983). A note on 'Locating a special population using random digit dialing'. *Public Opinion Quarterly*, 47, 576-578.
- WAKSBERG, J. (1984). *Efficiency of Alternative Methods of Establishing Cluster Sizes in RDD Sampling*. Memorandum de Westat non publié.
- WAKSBERG, J., BRICK, J.M., SHAPIRO, G., FLORES-CERVANTES, I. et BELL, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 713-718.
- WERKING, G., TUPEK, A. R., et CLAYTON, R. L. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics* 4, 349-362.
- WHITE, A. A. (1983). Response rate calculation in RDD telephone health surveys: current practices. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 277-282.
- WHITMORE, R. W., MASON, R. E. et HARTWELL, T. D. (1985). Use of geographically classified telephone directory lists in multi-mode surveys. *Journal of the American Statistical Association* 80, 842-844.
- WILSON, P., BLACKSHAW, N. et NORRIS P. (1988). An evaluation of telephone interviewing on the British Labour Force Survey. *Journal of Official Statistics* 4, 385-400.
- WINTER, D. L. S., et CLAYTON, R. L. (1990). Speech data entry: results of the first test of voice recognition for data collection. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 387-392.
- WISEMAN, F. (1972). Methodological bias in public opinion surveys. *Public Opinion Quarterly* 36, 105-108.
- XU, M., BATES, B.J. et SCHWEITZER, J.C. (1993). The impact of messages on survey participation in answering machine households. *Public Opinion Quarterly*, 57, 232-237.

- ROUQUETTE, C. (2000). La percée du téléphone portable et d'Internet. *INSEE Première* No. 200. INSEE, Paris. [http://www.insee.fr/fr/ffcd/docs_ffc/fp700.pdf].
- ST. CLAIR, J., et MUIR, J. (1997). Household adoption of digital technologies. *Year Book Australia* 1997. Canberra: Australian Bureau of Statistics.
- SALMON, C.T., et NICHOLS, J.S. (1983). The next birthday method of respondent selection. *Public Opinion Quarterly*, 47, 270-276.
- SCHUBERT, F., et PETSKA, T. (1993). Turning administrative systems into information systems. *Journal of Official Statistics*, 9, 109-119.
- SCHMIEDESKAMP, J. W. (1962). Reinterviews by telephone. *Journal of Marketing*, 26, 28-34.
- SEBOLD, J. (1988). Survey period length, unanswered numbers, and nonresponse in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 247-256.
- SHANKS, J.M. (1983). The current status of computer assisted telephone interviewing: recent progress and future prospects. *Sociological Methods and Research*, 12, 119-142.
- SHANKS, J.M., NICHOLS, W.T., II et FREEMAN, H.E. (1981). The California Disability Survey: design and execution of a computer-assisted telephone study. *Sociological Methods and Research*, 10, 123-140.
- SHAPIRO, G. M., BATTAGLIA, M. P., HOAGLIN, D. C., BUCKLEY, P. et MASSSEY, J. T. (1996). Geographical variation in within-household coverage of households with telephones in an RDD survey. *Proceedings of the Section on Survey Research Methods*, 491-496.
- SHURE, G.E., et MEEKER, R.J. (1978). A mini-computer system for multi-person computer-assisted telephone interviewing. *Behavior Methods and Instrumentation*, (Avril) 196-202.
- SMITH, C., et FRAZIER, E. L. (1993). Comparison of traditional and modified Waksberg. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 926-931.
- SQUIRE, P. (1988). Why the 1936 Literary Digest poll failed. *Public Opinion Quarterly*, 52, 125-133.
- STATISTICS CANADA (2000). *Selected Dwelling Characteristics and Household Equipment*. Income Statistics Division. [http://www.statcan.ca/english/Pgdb/People/Families/fam109b.htm].
- STOCK, J. S. (1962). How to improve samples based on telephone listings. *Journal of Advertising Research*, 2, 3, 50-51.
- STOKES, L., et YEH, M.-Y. (1988). Searching for causes of interviewer effects in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 357-376.
- SUDMAN, S. (1966). New uses of telephone methods in survey research. *Journal of Marketing Research*, 3, 107-120.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- SUDMAN, S. (1978). Optimum cluster designs within a primary unit interviewing. *Journal of the American Statistical Association*, 73, 300-304.
- SURVEY RESEARCH CENTER (2000). *Sample Design for Household Telephone Surveys: A Bibliography* 1949-1996. College Park, MD: University of Maryland. [http://www.bsos.umd.edu/src/sampbib.html].
- SURVEY SAMPLING INC. (1998). Random digit samples - part 1. [http://www.ssisamples.com/ssi.x20\$ssi.gen.search_item?id=119].
- STATISTICS NETHERLANDS (1987). *Automation in Survey Processing*. Voorburg/Heerlen: Netherlands Central Bureau of Statistics (CBS Select 4).
- SYKES, W.M., et COLLINS, M. (1987). Comparison entre l'interview téléphonique et l'interview sur place au Royaume-Uni. *Techniques d'enquête* 3, 19-33.
- SYKES, W.M., et COLLINS, M. (1988). Effects of mode of interview: experiments in the UK. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 301-320.
- THORNBERY, O.T. JR., et MASSSEY, J.T. (1978). Correcting for undercoverage bias in random digit dialed National Health Surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 224-229.
- THORNBERY, O.T. JR., et MASSSEY, J.T. (1983). Coverage and response in random digit dialed national surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 654-659.
- THORNBERY, O.T. JR., et MASSSEY, J.T. (1988). Trends in United States telephone coverage across time and subgroups. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 25-49.
- TORTORA, R.D. (1985). CATI in an agricultural statistics agency. *Journal of Official Statistics*, 1, 301-314.
- TOURANGEAU, R., et SMITH, T.W. (1998). Collecting sensitive information with different modes of data collection. *Computer Assisted Survey Information Collection*, (M.P. Cooper, et coll. - eds.). New York: John Wiley and Sons, 431-453.
- TRAUGOTT, M.W., GROVES, R.M. et LEFKOWSKI, J.M. (1987). Using dual frame designs to reduce nonresponse in telephone surveys. *Public Opinion Quarterly*, 51, 522-539.
- TREWIN, D., et LEE, G. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 3-24.
- TROLDAL, V.C., et CARTER, R.E. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research*, 1, 71-76.
- TUCKEL, P.S., et FEINBERG, B.M. (1991) The answering machine poses many questions for telephone survey researchers. *Public Opinion Quarterly*, 55, 200-217.
- TUCKEL, P.S., et O'NEILL, H. (1996). New technology and nonresponse bias in RDD surveys. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 889-894.
- TUCKEL, P., et SHUKERS, T. (1997). The effect of different introductions and answering machine messages on response rates. *Proceedings of the Section on Survey Research Methods*. American Statistical Association, 1047-1051.

- NORRIS, D.A., et PATON, D.G. (1991). L'Enquête sociale générale canadienne : bilan des cinq premières années. *Techniques d'enquête*, 17, 245-260.
- NTIA (2000). *Falling Through the Net, Toward Digital Inclusion*. Washington DC: National Telecommunications and Information Administration.
- NUSSEER, S., et THOMPSON, D. (1998). Web-based survey tools. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 951-956.
- OAKES, R.H. (1954). Differences in responsiveness in telephone versus personal interviews. *Journal of Marketing*, 19, 169.
- OKSENBURG, L., et CANNEL, C. (1988). Effects of interviewer vocal characteristics on nonresponse. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.), New York: John Wiley and Sons, 257-269.
- OLDENDICK, R.W., BISHOP, G.F., SORENSON, S.B. et TUCHFARBBER, A.J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics*, 4, 307-318.
- OLDENDICK R.W., et LINK, M.W. (1994). The answering machine generation: who are they and what problem do they pose for survey research? *Public Opinion Quarterly*, 58, 264-273.
- OFTEL (1999). *Homes Without a Fixed Line Phone - Who Are They?* [http://www.oftel.gov.uk/publications/research/unph0400.htm].
- OFTEL (2000). *Consumers' use of Internet*. [http://www.oftel.gov.uk/publications/research/int1000.htm]
- O'REILLY, J.M., HUBBARD, M.T., LESSLER, J.T., BIEMER, P.P., et TURNER, C.F. (1994). Audio and video computer assisted self-interviewing; preliminary tests of new technologies for data collection. *Journal of Official Statistics*, 10, 197-214.
- O'ROURKE, D., et BLAIR, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research*, 20, 428-432.
- PALIT, C.D. (1980). A microcomputer based computer assisted interviewing system. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 243-244.
- PALIT, C.D. (1983). Design strategies in RDD sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 627-629.
- PALIT, C.D., et BLAIR, J. (1986). Some alternatives for the treatment of first phase telephone numbers in a Waksberg-Mitofsky RDD sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 363-369.
- PALIT, C.D., et SHARP, H. (1983). Microcomputer-assisted telephone interviewing. *Sociological Methods and Research*, 12, 169-189.
- PANNEKOEK, J. (1988). Interviewer variance in a telephone survey. *Journal of Official Statistics*, 4, 375-384.
- PAYNE, S. L. (1956). Some advantages of telephone surveys. *Journal of Marketing*, 20, 278-281.
- PERNEGGER, T.V., MYERS, T.L., KLAG, M.J. et WHELTON, P.K. (1993). Effectiveness of the Waksberg telephone sampling method for the selection of population controls. *American Journal of Epidemiology*, 138, 574-584.
- PERONE, C., MATRUNDOLA, G. et SOVERINI, M. (1999). A quality control approach to mobile phone surveys; the experience of Telecom Italia Mobile. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 180-187.
- PERRY, J. B. (1968). A note on the use of telephone directories as a sample source. *Public Opinion Quarterly*, 32, 691-695.
- PHIPPS, P.A., et TUPEK, A.R. (1991). Fiabilité des données introduites au moyen d'un téléphone à clavier. *Techniques d'enquête*, 17, 17-30.
- PIAZZA, T. (1993). Meeting the challenge of answering machines. *Public Opinion Quarterly*, 57, 219-231.
- POTTER, F.J., MCNEILL, J.T., WILLIAMS, S.R. et WAITMAN, M.A. (1991). List-assisted RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 117-122.
- POTTHOFF, R.F. (1987a). Some generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- POTTHOFF, R.F. (1987b). Generalizations of the Mitofsky-Waksberg technique for random digit dialing: some added topics. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 615-620.
- POYNTER, R. (2000). *We've Got Five Years*. London: Association for Survey Computing meeting on Survey Research on the Internet.
- RAMOS, M., SEDIVJ, B.M. et SWEET, E.M. (1998). Computerized self-administered questionnaires. *Computer Assisted Survey Information Collection* (M.P. Couper, et coll. - eds.). New York: John Wiley and Sons, 389-408.
- RANTA-AHO, M., et LEPPINEN, A. (1997). Matching telecommunications services with user communication needs. *Proceedings of the International Symposium on Human Factors in Telecommunications*, (K. Nordby et L. Grafisk - eds.), Oslo, Norway, 401-408. [http://www.comlab.hut.fi/hft/publications/matcharticle.pdf].
- RICH, C.L. (1977). Is random digit dialing really necessary? *Journal of Marketing Research*, 14, 300-305.
- ROGERS, T.F. (1976). Interviews by telephone and in person: quality of responses and field performance. *Public Opinion Quarterly*, 40, 51-65.
- ROGERS, S.M., MILLER, H.G., FORSYTH, B.H., SMITH, T.K. et TURNER, C.F. (1996). Audio-CASI: the impact of operational characteristics on data quality. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1042-1047.
- ROMUALD, K.S., et HAGGARD, L.M. (1994). The effect of varying RDD telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1299-1304.
- ROSEN, R.J., MANNING, C.D. et HARRELL, L.J., Jr. (1998). Web-based data collection in the current employment statistics survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 354-359.
- ROSLOW, S., et ROSLOW, L. (1972). Unlisted phone subscribers are different. *Journal of Advertising Research*, 7, 4, 35-38.

- KUUSELA, V., et NOTKOLA, V. (1999). Survey quality and mobile phones. Article présenté à l'*International Conference on Survey Nonresponse*, Portland OR.
- KUUSELA, V., et VIRKI, K. (1999). Change of telephone coverage due to mobile phones. Article présenté à l'*International Conference on Survey Nonresponse*, Portland OR.
- LARSON, O. N. (1952). The comparative validity of telephone and face-to-face interviews in the measurement of message diffusion from leaflets. *American Sociological Review*, 17, 471-476.
- LAVRAKAS, P.J. (1993). *Telephone Survey Methods: Sampling, Selection and Supervision* (2^e édition). Newbury Park, CA: Sage Publications.
- LAVRAKAS, P.J., BAUMAN, S.L. et MERKLE, D.M. (1993). The last-birthday method and within-unit coverage problems. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1107-1112.
- LEPKOWSKI, J.M. (1988). Telephone sampling methods in the United States. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 73-98.
- LEPKOWSKI, J.M., et GROVES, R.M. (1984). The impact of bias on dual frame survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 265-270.
- LEPKOWSKI, J.M., et GROVES, R.M. (1986a). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 357-362.
- LEPKOWSKI, J.M., et GROVES, R.M. (1986b). A mean square error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.
- LEUTHOLD, D.A., et SCHEELE, R. (1971). Patterns of bias in samples based on telephone directories. *Public Opinion Quarterly*, 35, 249-257.
- LOCHANDER, W., SUDMAN, S. et BRADBURN, N. (1976). An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 71, 269-275.
- LYBERG, L., et KASPRZK, D. (1991). Data collection methods and measurement error: an overview. *Measurement Errors in Surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz et S. Sudman - éds.). New York: John Wiley and Sons, 237-258.
- MCCARTHY, W.F., et BATEMAN, D.V. (1988). The use of mathematical programming for designing dual frame surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 652-653.
- MCKAY, R.B., ROBISON, E.L. et MALIK, A.B. (1994). Touch-tone data entry for household surveys: research findings and possible applications. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 509-511.
- MAFFEO, C., FREY, W. et KALTON, G. (2000). Survey design and data collection issues in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.
- MAKLAN, D., et WAKSBERG, J. (1988). Within household coverage in RDD surveys. *Telephone Survey Methodology* (R. M. Groves, et coll. - éds.). New York: John Wiley and Sons, 51-69.
- MATLAKHOFF, L.A., et APPEL, M. V. (1997). The development of a voice recognition prototype for field listing. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 234-238.
- MASON, R.E., et IMBERMAN, F.W. (1988). Minimum cost sample allocation for Mitofsky-Waksberg random digit dialing. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 127-141.
- MASSEY, J.T. (1995). Estimating the response rate in a telephone survey with screening. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 673-677.
- MASSEY, J.T. et BOTMAN, S.L. (1988). Weighting adjustments for random digit dialed surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 143-160.
- MASSEY, J.T., O'CONNOR, D. et KROTKI, K. (1997). Response rates in random digit dialing (RDD) telephone surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 707-712.
- MEERKS, R.L., LANIER, A.T., FECISO, R.S. et COLLINS, M.A. (1998). Web-based data collection in national science foundation surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-353.
- MERKLE, D.M., BAUMAN, S.L. et LAVRAKAS, P.J. (1993). The impact of callbacks on survey estimates in an annual RDD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1070-1075.
- MITCHELL, G.H., et ROGERS, E.M. (1958). Telephone interviewing. *Journal of Farm Economics*, 40, 743-747.
- MITOFSKY, W. (1970). Sampling of Telephone Households. *Mémorandum de CBS non publié*.
- MOHADJER, L. (1988). Stratification of prefix areas for sampling rare populations. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 161-173.
- MULLET, G.M. (1982). The efficacy of plus-one dialing: self-reported status. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 575-576.
- NATHAN, G., et AFRAMIAN, N. (1996). An experiment with CATI in Israel. Article présenté à l'*InterCasic '96 Conference*, San Antonio, TX.
- NATHAN, G., et ELIAV, T. (1988). Comparison of measurement errors for telephone interviewing and home visits by misclassification models. *Journal of Official Statistics*, 4, 363-374.
- NICHOLS, W.L., II (1983). CATI research and development at the Census Bureau. *Sociological Methods and Research*, 12, 191-198.
- NICHOLS, W.L., II (1988). Computer-assisted telephone interviewing: A general introduction. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.). New York: John Wiley and Sons, 377-386.
- NICOLAAS, G., LYNN, P. et LOUND, C. (2000). Random digit dialling in the UK: viability of the sampling method revisited. Article présenté à *Fifth International Conference of Social Science Methodology*, Köln.

- GLASSER, G.J., et METZGER, G.D. (1972). Random digit dialing as a method of telephone sampling. *Journal of Marketing Research*, 9, 59-64.
- GLASSER, G.J., et METZGER, G.D. (1975). National estimates of nonlisted telephone households. *Journal of Marketing Research*, 12, 359-361.
- GOKSEL, H., JUDKINS, D.R. et MOSHER, W.D. (1991). Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 581-586.
- GROVES, R.M. (1977). *An Empirical Comparison of Two Telephone Sample Designs*. Rapport non publié Survey Research Center, the University of Michigan, Ann Arbor, MI.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II et WAKSBERG, J. - eds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- GROVES R.M., et KAHN, R.L. (1979). *Surveys by Telephone: A National Comparison With Personal Interview*. New York: Academic Press.
- GROVES, R.M., et LEFKOWSKI, J.M. (1985). Dual frame, mixed mode survey designs. *Journal of Official Statistics*, 1, 264-286.
- GROVES, R.M., et LEFKOWSKI, J.M. (1986). An experimental implementation of a dual frame telephone sample design. *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 340-345.
- GROVES, R.M., et LYBERG, L.E. (1988a). An overview of nonresponse issues in telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 191-211.
- GROVES, R.M., and LYBERG, L.E. - Eds. (1988b). Telephone survey methodology. *Journal of Official Statistics* (numéro spéciale), 4, 283-416.
- GUNN, W.J., et RHODES, I.N. (1981). Physician response rates to a telephone survey: effects of monetary incentive level. *Public Opinion Quarterly*, 45, 109-115.
- HAGAN, D. E., et MEIER C. C. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly*, 47, 547-556.
- HARLOW, B.L., CREA, E.C., EAST, M.A., OLESON, B., FRAER, C.J. et CRAMER, D.W. (1993). Telephone answering machines: the influence of leaving messages on telephone interviewing response rates. *Journal of Epidemiology*, 4, 380-383.
- HARTGE, P., BRINTON, L.A., ROSENTHAL, J.F., CAHILL, J.L., HOOVER, R.N. et WAKSBERG, J. (1984). Random digit dialing in selecting a population-based control group. *American Journal of Epidemiology*, 120, 825-833.
- HAUCK, M., et COX, M. (1974). Locating a sample by random digit dialing. *Public Opinion Quarterly*, 38, 253-260.
- HERZOG, A.R., et RODGERS, W.L. (1988). Interviewing older adults: mode comparison using data from a face-to-face survey and a telephone resurvey. *Public Opinion Quarterly*, 52, 84-99.
- HOAGLIN, D.C., et BATTAGLIA, M.P. (1996). A comparison of two methods of adjusting for noncoverage of nontelephone households in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 497-502.
- HOCHSTETM, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.
- HOGUE, C.R., et CHAPMAN, D.W. (1984). An investigation of PSU cutoff points for a random digit dialing survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 286-291.
- HOX, J.J., DE LEEUW, E.D. et KREFT, J.G.G. (1991). The effect of interviewer and respondent characteristics on the quality of telephone data: a multilevel model. *Measurement errors in surveys* (P.P. Biemer, L.E. Lyberg, N.A. Mathiowetz et S. Sudman - eds.). New York: John Wiley and Sons, 439-461.
- INGLIS, K.M., GROVES, R.M. et HERRINGA, S.G. (1987). Plans de sondage d'enquêtes téléphoniques auprès de ménages noirs aux États-Unis. *Techniques d'enquête*, 13, 1-17.
- JANOFSKY, A.L. (1971). Affective self-disclosure in telephone versus face-to-face interviews. *Journal of Humanistic Psychology*, 11, 93-103.
- JOHNSON, T., FENDRICH, M., SHALIGRAM, C. et GAREY, A. (1997). A comparison of interviewer effects models in an RDD telephone survey of drug use. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 887-892.
- KATLSBEER, W.D., et DURHAM, T.A. (1994). Nonresponse and its effects in a followup telephone survey of low-income women. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 943-948.
- KALTON, G. (2000). L'évolution de la recherche sur les enquêtes au cours des 25 dernières années. *Techniques d'enquête*, 26, 3-11.
- KATZ, D., et CANTRIL, H. (1937). Public opinion polls. *Sociometry*, 1, 155-179.
- KATZ, J.E., et ASPDEN, P. (1998). Internet dropouts in the USA. *Telecommunications Policy*, 22, 4/5, 327-339.
- KEETTER, S. (1995). Estimating telephone noncoverage bias with a telephone survey. *Public Opinion Quarterly*, 59, 196-217.
- KEHOE, C., PITKOW, J., SUTTON, K., AGGARWAL, G. et ROGERS, J.D. (1999). *Results of GVU's Tenth World Wide Web User Survey*. Atlanta, GA: Graphics Visualization and Usability Center, College of Computing, Georgia Institute of Technology. [http://www.gvu.gatech.edu/user_surveys].
- KHURSHID, A., et SAHAI, H. (1995). A bibliography on telephone survey methodology. *Journal of Official Statistics*, 11, 325-367.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KOEPSSEL, T.D., MCGUIRE, V., LONGSTRETH, Jr., W.T., NELSON, L.M. et VAN BELLE, G. (1996). Randomized trial of leaving messages on telephone answering machines for control recruitment in an epidemiological study. *American Journal of Epidemiology*, 144, 704-706.
- KORMENDI, E. (1988). The quality of income information in telephone and face to face surveys. *Telephone Survey Methodology* (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 341-356.

- CZAJA, R., BLAIR, J. et SEBESTIK, J.P. (1982). Respondent selection in a telephone survey: a comparison of three techniques, *Journal of Marketing Research*, 19, 381-385.
- DECISION ANALYST (1997). 'More households using answering machines. *News Release*, October 15, 1997. [http://www.decisionanalyst.com/publ_data/1997/ansmach1.htm].
- DEKKER, F., et DORN, P.K. (1984). Computer Assisted Telephone Interviewing: A research project in the Netherlands. Article présenté à la *Conference of the Institute of British Geographers*. DE LEEUW, E.D., et VAN DER ZOOEWEN, J. (1988). Data quality in telephone and face to face surveys: a comparative meta-analysis. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.), New York: John Wiley and Sons, 283-299.
- DIEHR, P., KOEPEL, T.D., CHEADLE, A. et PSATY, B.M. (1992). Assessing response bias in random-digit dialing surveys: The telephone-prefix method. *Statistics in Medicine*, 11, 1009-1021.
- DILTMAN, D.A. (1978). *Mail and Telephone Surveys: the Total Design Method*. New York: John Wiley and Sons.
- DREW, J.D., CHOUDHRY, G.H. et HUNTER, L.A. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.), New York: John Wiley and Sons, 233-246.
- DREW, J.H., et GROVES, R.M. (1989). Adjusting for nonresponse in a telephone subscriber survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 452-456.
- DUTKA, S., et FRANKEL, L. R. (1980). Sequential survey design through the use of computer assisted telephone interviewing. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 73-76.
- EASTLACK, I.O., JR., et ASSAEL, H. (1966). Better telephone surveys through centralized interviewing. *Journal of Advertising Research*, 6, 1, 2-7.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1984). *The Role of Telephone Data Collection in Federal Statistical Policy Working Paper 12*, Washington, D.C.
- FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY. (1990). *Computer Assisted Survey Information Collection*, Statistical Policy Working Paper 19, Washington, D.C.
- FEDERAL REPUBLIC OF GERMANY (1999). Continuous family budget surveys for January 1999. Statistisches Bundesamt: *Press release*, 20 December, 1999. [http://www.statistik-bund.de/press/enGLISH/p4350024.htm].
- FELSON, L. (2001). Netting limitations: online researchers' new tactics for tough audiences. *Marketing News* (American Marketing Association), 35, 5 [http://www.ama.org/pubs/article.asp?id=4881].
- PINK, J.C. (1983). CATI's first decade: The Chilton experience. *Sociological Methods and Research*, 12, 153-168.
- FISCHBACHER, C., CHAPPEL, D., EDWARDS, R. et SUMMERTON, N. (1999). The use and abuse of the Internet for survey research. *Proceedings of the Association for Survey Computing 3rd International Conference*, Edinburgh, 501-507.
- FITZ, J.E. (1979). Some results from the telephone health interview system. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 244-249.
- FOLEY, D.J., et BROCK, D.B. (1990). Comparison of in-person and telephone responses in a survey of the last days of life. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 382-386.
- FORD, E.S. (1998). Characteristics of survey participants with and without a telephone: findings from the third National Health and Nutrition Examination Survey. *Journal of Clinical Epidemiology*, 51, 55-60.
- FORSMAN, G. (1993). Sampling individuals within households in telephone surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1113-1118.
- FORSMAN, G., et DANIELSSON, S. (1997). Can plus digit sampling generate a probability sample? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 958-963.
- FOX, A., et RILEY, J. P. (1996). Telephone coverage, housing quality and rents: RDD survey biases. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 515-519.
- FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. et SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 934-937.
- FREEEMAN, H.E., et SHANKS, J. M. - Eds. (1983). The emergence of computer-assisted survey research. *Sociological Methods and Research*, 12 (numéro spécial), 115-230.
- FRÉJEAN, M., PANZANI, J.-P. et TASSI, P. (1990). Les ménages inscrits en liste rouge et les enquêtes par téléphone. *Journal de la Société de Statistique de Paris*, 131, Nos. 3-4, 86-102.
- FRÉY J. H. (1989). *Survey Research by Telephone* (2^e édition). Beverly Hills, CA: Sage Publications.
- FRY, H.G., et MCNAIRE, S. (1958). Data gathering by long distance telephone. *Public Health Records*, 73, 831-835.
- GABLER, S., et HAEDER, S. (2000). 'Telephone sampling in Germany. *Paper presented at Fifth International Conference of Social Science Methodology*, Köln.
- GENESYS (1996). Unlisted numbers: what's really important. *Genesys News* (Genesys Sampling Systems, Fort Washington, PA), 1-2.
- GHOSH, D. (1984). Improving the plus 1 method of random digit dialing. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 285-288.
- GIESBRECHT, L.H., KULP, D.W. et STARER, A.W. (1996). Estimating coverage bias in RDD samples with current population survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 503-508.

- BRUNNER, J.A., et BRUNNER, G.A. (1971). Are voluntary unlisted telephone subscribers really different? *Journal of Marketing Research*, 8, 121-124.
- BYRANT, B.E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, 12, 129-135.
- BYRON, M.C. (1976). The literary digest poll: Making of a statistical myth. *The American Statistician*, 30, 184-185.
- BULL, S.B., PEDERSON, L.T., et ASHLEY, M.J. (1988). Intensity of follow up: effects on estimates in a population telephone survey with an extension of Kish's (1965) approach. *American Journal of Epidemiology*, 127, 552-561.
- BURKE, J., MORGANSTEIN, D., et SCHWARTZ, S. (1981). Toward the design of an optimal telephone sample. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 448-453.
- CAHALAN, D. (1960). Measuring newspaper readership by telephone: two comparisons with face to face interviews. *Journal of Advertising Research*, 1, 2, 1-6.
- CAHALAN, D. (1989). Comment: The digest poll rides again! *Public Opinion Quarterly*, 53, 129-133.
- CAMPBELL, J., et PALIT, C.D. (1988). Total digit dialing for a small area census by phone. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 549-551.
- CANNELL, C.F., GROVES, R.M., MAGILAVY, L.J., MATHIOWETZ, N.A., MILLER, P.V., et THORNBERRY, O.T. (1987). An experimental comparison of telephone and personal health interview surveys. *Vital and Health Statistics, Series 2*, 106, Public Health service.
- CARON, P., et LAVALLÉE, P. (1998). Comparison study on the quality of financial data collected through personal and telephone interviews. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 208-213.
- CASADY, R.J., et LEPKOWSKI, J.M. (1991). Optimal allocation for stratified telephone survey design. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 111-116.
- CASADY, R.J., et LEPKOWSKI, J.M. (1993). Plans d'enquête téléphonique stratifiés. *Techniques d'enquête*, 19, 115-125.
- CASADY, R.J., et LEPKOWSKI, J.M. (1998). Telephone sampling. *Encyclopedia of Biostatistics*. New York: John Wiley and Sons, 4498-4511.
- CASADY, R.J., et SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-609.
- CASRO (1982). *Report of the Council of American Survey Research Organization Completion Rate Task Force*. New York: Audits and Surveys Inc. (Rapport non-publié).
- CENTRAL BUREAU OF STATISTICS (2000). *The Household Expenditure Survey 1999*. Special Publication 1147. Jerusalem.
- CUMMINGS, M.K. (1979). Random digit dialing: a sampling technique for telephone surveys. *Public Opinion Quarterly*, 43, 233-244.
- CHAPMAN, D.W., et ROMAN, A.M. (1985). An investigation of substitution for an RPD survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 269-274.
- CHOUDHRY, G.H. (1989). Cost-variable optimization of dual frame design for estimating proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 566-571.
- CLAYTON, R.L., et WINTER, D.L.S. (1992). Speech data entry: results of a test of voice recognition for survey data collection. *Journal of Official Statistics*, 8, 377-388.
- COLLINS, M. (1983). Computer assisted telephone interviewing in the UK. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 636-639.
- COLLINS, M. (1999). Editorial: sampling for UK telephone surveys. *Journal of the Royal Statistical Society, A* 162, 1-4.
- COLLINS, M., SYKES, W., WILSON, P., et BLACKSHAW, N. (1988). Nonresponse: the UK experience. *Telephone Survey Methodology*, (R.M. Groves, et coll. - eds.). New York: John Wiley and Sons, 213-232.
- COLOMOTOS, J. (1965). The effects of personal vs. telephone interviews on socially acceptable responses. *Public Opinion Quarterly*, 29, 457-458.
- COOMBS, L., et FREEDMAN, R. (1964). Use of telephone interviews in a longitudinal fertility study. *Public Opinion Quarterly*, 28, 112-117.
- COOPER, S.L. (1964). Random sampling by telephone: an improved method. *Journal of Marketing Research*, 1, 45-48.
- COOPER, M.P., BAKER, R.P., BETHLEHEM, J., CLARK, C.Z.F., MARTIN, J., NICHOLS, W.L., II et O'REILLY, J.M. - (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley and Sons.
- COOPER, M.P., BLAIR, J., et TRIPLETT, T. (1999). A comparison of mail and e-mail for a survey of employees in US statistical agencies. *Journal of Official Statistics*, 15, 39-56.
- COOPER, M.P., et NICHOLS, W.L., II (1998). The history and development of computer assisted survey information collection methods. *Computer Assisted Survey Information Collection*, (M.P. Couper, et coll. - eds.). New York: John Wiley and Sons, 1-22.
- CUNNINGHAM, P., BRICK, J.M., et MEADER, J. (2000). 1999 *NSAF In-Person Survey Methods Report No. 5*. Washington, DC: Urban Institute. [http://newfederalism.urban.org/nsaf/methodology_pjrs/1999/Methodology_5.pdf].

- BENNETT, C.T. (1961). A telephone interview: A method for conducting a follow-up study. *Mental Hygiene*, 45, 216-220.
- BERGINI, D.H., et MASSIEY, J.T. (1979). Obtaining the household roster in a telephone survey: The impact of names and placement on response rates. *Proceedings of the Social Statistics Section, American Statistical Association*, 136-140.
- BERRY, S.H., et O'Rourke, D. (1998). Administrative designs for centralized telephone survey centers: Implications of the transition to CATI. *Telephone Survey Methodology*, (R.M. Groves, et coll. - éds.), New York: John Wiley and Sons, 457-474.
- BIEL, A.L. (1967). Abuses of survey research techniques: the phony interview. *Public Opinion Quarterly*, 31, 298.
- BIEMER, P.P. (1983). Optimal dual frame sample design: Results of a simulation study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 630-635.
- BINSON, D., CANGHOLA, J.A. et CATANIA, J.A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics*, 16, 53-59.
- BLAIR, J., et CZAJA, R. (1982). Locating a special population using random digit dialing. *Public Opinion Quarterly*, 46, 585-590.
- BLANKENSHIP, A.B. (1977a). *Professional Telephone Surveys*. New York: McGraw Hill.
- BLANKENSHIP, A.B. (1977b). Listed versus unlisted numbers in telephone-survey samples. *Journal of Advertising Research*, 39-42.
- BOTMAN, S.L., et ALLEN, K. (1990). Some effects of undercoverage in a telephone survey of teenagers. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-400.
- BOTMAN, S.L., MASSIEY, J.T. et SHIMIZU, I.M. (1982). Effect of weighting adjustments on estimates from a random-digit-dialed telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 139-144.
- BRICK, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methodology, American Statistical Association*, 296-301.
- BRICK, J.M., et COLLINS, M.A. (1997). A response rate experiment for RDD surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1052-1057.
- BRICK, J.M., KATTON, G., NIXON, M., GIVENS, J. et EZZATI-RICE, T. (2000). Statistical issues in a record check study of childhood immunizations. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference: Statistical Policy Working Paper*, 30, 625-634.
- BRICK, J.M., et WAKSBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Techniques d'enquête*, 17, 31-46.
- BRICK, J.M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.
- BRICK, J.M., WAKSBERG, J., KULP, D. et STARKER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.
- BATTAGLIA, M.P., SHAPIRO, G. et ZELL, E.R. (1996). Substantial response bias may remain when records are used in a telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 452-455.
- BANKS, R., CHRISTIE, C., CURRALL, J., FRANCIS, J., HARRIS, P., LEE, B., MARTIN, J., PAYNE, C. et WESTLAKE, A. (éds.) (1999). ASC'99 - Leading Survey & Statistical Computing into the New Millennium. *Proceedings of the ASC International Conference*. Association for Survey Computing Chesham, Bucks, UK.
- BANKS, M.J., et HAGAN, D.E. (1984). Reducing interviewer screening and controlling sample size in a local-area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 271-273.
- BAKER, R.P. (1998). The CASIC future. *Computer Assisted Survey Information Collection*, (M.P. Couper et coll. - éds.). New York: John Wiley and Sons, 583-604.
- AQUILINO, W.S., et LO SCUTO, L.A. (1990). Effects of interview mode on self-reported drug use. *Public Opinion Quarterly*, 54, 362-395.
- ANDERSON, J.E., NELSON, D.E. et WILSON, R.W. (1998). Telephone coverage and measurement of health risk indicators: data from the National Health Interview Survey. *American Journal of Public Health*, 88, 1392-1395.
- AMERICAN STATISTICAL ASSOCIATION (1999). *More About Telephone Surveys*. ASA series: what is a survey? Section on Survey Research Methods [http://www.amstat.org/sections/srms/brochures/telephone.pdf].
- ALEXANDER, C.H. (1988). Cutoff rules for secondary calling in a random digit dialing survey. *Telephone Survey Methodology*, (R.M. Groves et coll. - éds.). New York: John Wiley and Sons, 113-126.
- ABRAHAM, S.Y., STEIGER, D.M. et SULLIVAN, C. (1998). Electronic and mail self-administered questionnaires : A comparative assessment of use among elite populations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 833-841.
- POUR conclure, la méthodologie des téléenquêtes, dont l'évolution au cours des dernières décennies a rendu les enquêtes téléphoniques viables et en a fait l'instrument prédominant d'enquête, devra être mise à jour continuellement en vue de s'adapter à l'évolution constante des technologies de télécommunication et de leur utilisation. Cependant, les éléments méthodologiques fondamentaux de cette évolution existent et continueront de permettre d'utiliser des solutions de pointe pour obtenir des données d'enquête de haute qualité.

BIBLIOGRAPHIE

devraient être recueillis auprès des membres des ménages et comprendre ceux sur la taille du ménage pour que l'on puisse pondérer correctement les données sur les caractéristiques du ménage. Si le système de numéros de communication permet d'attribuer un numéro unique à chaque personne, aucun renseignement ne sera nécessaire sur les modes de communication ou leur multiplicité.

S'il n'existe pas de listes de numéros de communication ou que le problème des numéros non publiés persiste, il faudra recourir à une forme de sélection par CA. La méthode suivie ne devrait pas différer beaucoup des méthodes de CA employées à l'heure actuelle. Si l'on suppose que le système de numéros de communication sera effectivement unique et universel et qu'il suivra une certaine logique, on devrait pouvoir élaborer facilement des méthodes efficaces d'échantillonnage. Idéalement, le système de numérotation présentera certains liens avec les données géographiques, grâce à l'adresse permanente de la personne. Sinon, les enquêtes locales, voire même nationales, par CA deviendront extrêmement difficiles à concevoir efficacement. L'obtention de suffisamment de renseignements sur le système de numérotation permettrait de réduire au minimum le nombre de numéros hors du champ de l'enquête.

Puisque le répondant contrôlera vraisemblablement en grande partie le choix du mode de communication, la question de la répartition des unités d'échantillonnage entre les modes de communication ne se posera pour ainsi dire pas. Les concepteurs d'enquête devront préparer toute une gamme d'instruments de collecte adaptés aux divers modes de communication. Cette gamme inclura des instruments textuels, comme des versions pour télécopieur, courrier électronique et Internet du questionnaire, des instruments oraux, comme les interviews orales classiques et automatisées ou une combinaison de celles-ci. L'intégration des données recueillies selon divers modes de collecte en un ensemble uniforme de données posera un défi technologique énorme, mais surmontable.

La situation presque utopique susmentionnée ne se concrétisera sans doute pas avant longtemps et, entre-temps, il faudra élaborer des méthodes appropriées pour résoudre les problèmes que posera, à court terme, la mise au point des technologies de communication et de leurs applications. Comme il l'est noté à la section 4.3, il faudra tenir compte prochainement de la nécessité de passer des enquêtes téléphoniques fondées uniquement sur le service téléphonique fixe à une combinaison de services mobile et fixe. Fondamentalement, la méthodologie des bases de sondages multiples mise au point pour couvrir à la fois les ménages qui ont le téléphone et ceux qui ne l'ont pas peut être étendue facilement pour résoudre ce problème. Il faut encore créer les bases de sondages et/ou les méthodes d'échantillonnage par CA appropriées pour les téléphones mobiles, mais les principes fondamentaux existent. Il faudra aussi résoudre le problème que pose la combinaison de données obtenues à partir de téléphones mobiles, qui sont

télécommunications, sera sans doute celui du choix d'un cadre de référence pertinent et de la répartition des unités d'échantillonnage entre les modes de collecte. On pense qu'en bout de ligne, chaque personne se verra attribuer un numéro de communication personnel (ou numéro d'identification) unique permanent grâce auquel elle pourra être rejointe de diverses façons (communication écrite, verbale ou visuelle) grâce à divers appareils fixes ou sans fil qui pourraient se trouver à son domicile, à son bureau ou être mobiles. Le choix du mode d'interview dépendra de la décision collective du répondant et de l'appelant. Les libératoires condamneront indubitablement l'idée de ce numéro universel (qui serait essentiellement un numéro d'identité), mais il est probable qu'il devienne éventuellement acceptable, même si de petits groupes d'activistes essayent d'éviter de l'utiliser et même d'entraver son application. En fait, des systèmes normalisés de numéros d'identité universels sont en place et bien acceptés depuis plusieurs décennies dans plusieurs pays d'Europe du Nord et en Israël. Dans ces pays, le numéro d'identité, qui n'est pas considéré comme un renseignement confidentiel, est utilisé à grande échelle à de nombreuses fins administratives et commerciales. Par exemple, en Israël, la loi exige que les chèques personnels portent le numéro d'identité-cation de la personne, son nom, son adresse et son numéro de téléphone.

Une fois que l'on pourra exploiter un système de numéro de communication unique, on pourra appliquer les méthodes types d'échantillonnage. Il se pourrait fort bien que l'accès à des listes complètes de ces numéros – ne contenant peut-être que des renseignements géographiques et autres limités, soit généralisé, comme pour les numéros d'identification utilisés dans de nombreux systèmes nationaux d'enregistrement. Il y a de bonnes raisons de croire que la situation sera la même pour les numéros de communication – au départ, au moins en Europe, plutôt qu'en Amérique du Nord. En effet, les techniques perfectionnées de filtrage pourraient fort bien rendre redondante l'utilisation de numéros de téléphone non publiés. Le filtrage risqué, certes, de faciliter la non-réponse, mais la possibilité de transmettre un message écrit par courrier électronique ou de laisser un message dans une boîte vocale pourrait réduire le problème.

L'échantillonnage à partir du genre de listes susmentionnées serait simple, mais inefficace dans la plupart des cas, puisque les données auxiliaires n'offriraient peut-être qu'un avantage marginal. Bien que les listes puissent permettre de faire la distinction entre les numéros résidentiels et commerciaux, il est douteux qu'elles fournissent des renseignements sur les ménages. Il s'ensuit que l'unité d'échantillonnage et de déclaration devrait être la personne plutôt que le ménage. Il s'agit de toute façon du but poursuivi par de nombreuses enquêtes et il n'est pas certain que le choix du ménage comme unité d'échantillonnage pour les téléenquêtes soit utile, même dans la pratique courante. S'ils sont nécessaires, les renseignements sur les ménages

d'une opération d'établissement de listes par les employés qui travaillent sur le terrain. Il convient de souligner que, si la SDC est de toute évidence propre aux enquêtes téléphoniques, la RVI peut être appliquée à d'autres modes de collecte.

Les méthodes d'auto-interview assistée par ordinateur (AIAO), qui incluent les modes audio (AIAOA) et vidéo (AIAOV) de collecte, sont considérées depuis longtemps comme des extensions naturelles des enquêtes par la poste qui tirent parti de la technologie moderne (Dillman 2000). On a surtout insisté sur leur utilité pour les enquêtes portant sur des sujets délicats et embarrassants, où l'idée que l'intervieweur soit présent (interview sur place) peut faire hésiter les répondants à participer à l'enquête. Pour une revue des progrès récents visant ces méthodes, consulter Baker (1998), O'Reilly, Hubbard, Lessler, Biemer et Turner (1994), Rogers, Smith et Turner (1996), ainsi que Tourangeau et Smith (1998). Presque toutes les applications décrites correspondent à des enquêtes pour lesquelles l'instrument est amené au domicile du répondant par un membre du personnel sur le terrain. Certains chercheurs ont déjà essayé d'utiliser le téléphone pour la collecte de données par AIAOA (AIAOA-T) – consulter Turner, Forsyth, O'Reilly, Cooley, Smith, Rogers et Miller (1998). La mise au point longtemps attendue de la vidéotéléphonie pour en faire une forme commune généralisée de service téléphonique pour les ménages ne s'est pas encore concrétisée. Le jour où elle se matérialisera, la vidéotéléphonie rendra possible l'AIAOV téléphonique (AIAOV-T), ce qui aura des conséquences importantes pour le travail d'enquête. L'ajout d'un élément visuel permettra de surmonter plus facilement les problèmes que posent aujourd'hui les enquêtes téléphoniques contrairement aux enquêtes par interview sur place (regard droit dans les yeux de l'intervieweur, utilisation de cartons aide-mémoire ou d'autres aides visuelles). Il faudra sans doute fort longtemps avant que la vidéotéléphonie soit universelle, si bien que, du moins pour le moment, l'AIAOV-T ne pourra servir que de moyen complémentaire de collecte des données.

4.3 Téléphones mobiles

Les problèmes de couverture que pourraient poser les enquêtes par CA axées sur le service téléphonique fixe à cause de la prolifération rapide des téléphones mobiles ont été mentionnés à la section 3.3.1. Dans l'avenir, il est évident qu'il faudra se servir des téléphones mobiles pour rejoindre le nombre sans cesse croissant de ménages qui ne sont pas abonnés au service téléphonique fixe. Le niveau actuel de couverture des téléphones mobiles signifie que les enquêtes par téléphone mobile ne peuvent, en général, être utilisées que pour des populations particulières ou pour compléter les enquêtes par CA dans le cas d'un service fixe. Par exemple, Perone, Matmundola et Sovetini (1999) présentent une enquête par téléphone mobile auprès d'une

« choisis de participer » plutôt qu'auprès d'un échantillon probabiliste.

Par ailleurs, il existe des preuves que la collecte de données par Internet donne d'assez bons résultats pour les enquêtes auprès des établissements. Nasser et Thompson (1998) décrivent son utilisation pour les National Resources Inventory Surveys du US Department of Agriculture; Rosen, Manning et Harrell (1998) publient les résultats de la collecte de données par Internet auprès des établissements universitaires, d'organismes fédéraux et de sociétés privées pour les US National Science Foundation surveys. En supposant que l'on arrivera à résoudre les problèmes de couverture et d'échantillonnage en ce qui concerne les ménages et les particuliers, il est permis d'espérer que l'on pourra appliquer la collecte de données par Internet aux enquêtes-ménages dans l'avenir.

4.2 Autres questionnaires à remplir soi-même informatisés (QRSM) et méthodes d'auto-interview assistée par ordinateur (AIAO)

Couper et Nichols (1998) font une distinction entre la collecte au moyen d'un questionnaire à remplir soi-même informatisé (QRSM) qui ne comporte pas l'intervention d'un intervieweur et l'auto-interview assistée par ordinateur (AIAO) durant laquelle un intervieweur est présent ou délivre l'instrument d'enquête. Donc, tant les enquêtes par courrier électronique que celles par Internet se fondent sur l'utilisation d'un QRSM appuyé par la technologie des télécommunications. La saisie des données au clavier (SDC), où les répondants entrent les données en se servant du clavier de leur téléphone, et la reconnaissance vocale interactive (RVI) ou la saisie par reconnaissance vocale (SRV) sont d'autres méthodes basées sur un QRSM. Dans les deux cas, les répondants font eux-mêmes l'appel à leur meilleure convenance pour faire leur déclaration, après une première prise de contact et les deux méthodes ont été mises à l'essai à grande échelle et utilisées avec succès par le US Bureau of Labor Statistics pour recueillir des données auprès des établissements pour son Current Employment Statistics program Working, Tupek et Clayton (1988), Winter et Clayton (1990), Clayton et Winter (1992), Phipps et Tupek (1991) décrivent une étude de la qualité de la collecte de données par SDC réalisée par vérification des enregistrements, et montrent que la méthode pose peu de problèmes et que les erreurs diminuent avec l'expérience. Plus récemment, les bureaux américains de la statistique ont entrepris des essais en vue de déterminer si l'est possible d'appliquer ces méthodes axées sur un QRSM aux enquêtes-ménages. McKay, Robison et Malik (1994) font le compte rendu d'un essai de laboratoire préliminaire de la SDC pour la Current Population Survey. Malakoff et Appel (1997), quant à eux, décrivent la mise au point d'un prototype de RVI au US Bureau of Census, mais il s'agit

Lors d'essais sur le terrain de la US National Study of Postsecondary Faculty de 1999, on a donné aux administrateurs ainsi qu'aux professeurs le choix de remplir et de renvoyer par la poste un questionnaire papier traditionnel ou d'utiliser un questionnaire à remplir soi-même informatisé (QRSMI) par Internet (Abraham, Steiger et Sullivan 1998). Bien qu'il soit raisonnable de supposer que presque tous les répondants avaient accès à Internet, 8 % seulement des professeurs répondants et 17 % des administrateurs d'établissements ont choisi le QRSMI. La US National Science Foundation prévoit offrir une option Internet pour la National Survey of Recent College Graduates de 1999, l'hypothèse étant que la plupart des personnes visées par l'enquête connaissent les ordinateurs et ont accès à Internet (Meeks, Lanier, Fecso et Collins 1998). Pour une revue de l'utilisation des QRSMI par les organismes gouvernementaux et les organismes privés d'enquête et des problèmes qu'elle pose, consulter Ramos, Sedivi et Sweet (1998).

Cependant, à l'heure actuelle, la plupart des enquêtes Internet visant des populations générales s'appuient sur l'échantillonnage non probabiliste, principalement par recours à une forme ou l'autre d'auto-sélection. Fischbacher, Chapel, Edwards et Summerton (1999) décrivent la méta-analyse de 28 enquêtes touchant le domaine de la santé réalisées par courrier électronique et par Internet. Nombre de ces enquêtes étant des études épidémiologiques visant des personnes atteintes de maladies particulières, le problème du biais de sélection empêche la généralisation de la plupart des résultats. L'une des enquêtes Internet les plus importantes est la WWW User Survey réalisée par le Graphics Visualization and Usability Center du Georgia Institute of Technology (Kehoe, Perkow, Sutton, Aggarwal et Rogers 1999). Bien que la population étudiée soit, par définition, les internautes, l'absence de tout cadre d'échantillonnage pour cette population sous-entend que les répondants ont dû être sollicités par diverses méthodes (annonces sur Internet et d'autres supports, bandeaux publicitaires, prix d'encouragement en espèces, etc.) au lieu d'être échantillonnés avec une probabilité connue. Bien que quelque 20 000 internautes aient participé, les auteurs du rapport mentionnent que les données sont biaisées en faveur des utilisateurs chevronnés et les plus fréquents d'Internet et recommandent que l'on augmente les données en réalisant des enquêtes auprès d'échantillons aléatoires. Pour essayer d'éviter le biais dû à la réalisation des enquêtes auprès d'échantillons de la population de personnes qui ont accès à Internet uniquement, certains organismes commerciaux d'enquête distribuent à tous les membres des panels qu'elles sélectionnent par CA un appareil qui leur permet de se raccorder à Internet au moyen de leur téléviseur pour s'assurer d'obtenir des résultats cohérents (Felson 2001). Cependant, Poynter (2000) prédit que d'ici à l'an 2005, 95 % d'études de marché seront réalisées par la voie d'Internet, mais que 80 % seront réalisées auprès de répondants qui auront

Canada, 22 % en Finlande, 7 % en France et 5 % en Belgique en 1999 selon Rouquette (2000), 12 % en Israël, (en 1999 – Bureau central de la statistique 2000) et 11 % en Allemagne (République fédérale d'Allemagne 1999). Malgré cette augmentation rapide, la couverture Internet est encore loin d'être complète. Qui plus est, certaines données laissent entendre qu'une catégorie croissante d'anciens internautes se forme parallèlement à l'augmentation globale de l'utilisation. Selon Katz et Aspdén (1998), la proportion d'anciens utilisateurs d'Internet est passée de 8 % en 1995 à 11 % en 1996. Cependant, l'augmentation globale de l'accès a favorisé l'utilisation du courrier électronique et de l'Internet pour la réalisation des enquêtes. Quoique la couverture d'une enquête par courrier électronique (ECE) soit comparable à celle d'une enquête Internet et que toutes deux se fondent sur l'utilisation d'un questionnaire à remplir soi-même informatisé (ou QRSMI), il existe une différence fondamentale entre ces deux formes de télé-enquêtes. L'enquête par courrier électronique est fort semblable à une enquête par la poste, en ce sens qu'elle consiste à envoyer un questionnaire textuel et à demander au répondant de le retourner dument rempli. L'avantage par rapport à l'enquête par la poste tient au coût plus faible, ainsi qu'à la facilité et à la simplicité de transmission et de réception. L'enquête Internet se fonde, en général, sur l'interaction entre le répondant et l'instrument d'enquête, grâce à l'utilisation de Java, XML ou d'un instrument comparable. Elle permet de nombreuses améliorations, comme l'utilisation de couleurs et d'animation, et les possibilités qu'elle offre sont multiples en ce qui concerne l'utilisation d'enchaînements complexes des questions et la vérification en temps réel. Les perspectives intéressantes de développement de systèmes novateurs de collecte des données grâce à la création incessante de nouveaux outils Internet ne permettent pas encore de surmonter le problème fondamental que posent les enquêtes par courrier électronique et les enquêtes Internet, c'est-à-dire la couverture absolument insuffisante à l'heure actuelle pour la plupart des populations humaines étudiées (Dillman 2000).

Néanmoins, les enquêtes par courrier électronique et par Internet peuvent et sont utilisées, avec plus ou moins de bonheur, pour certaines populations pour lesquelles la couverture est virtuellement complète, ou conjuguées à d'autres modes de collecte. Ainsi, Couper, Blair et Triplet (1999) font le compte rendu d'une étude expérimentale visant à comparer l'utilisation du courrier électronique et du service postal ordinaire pour réaliser une enquête auprès des employés de plusieurs bureaux gouvernementaux de la statistique aux États-Unis. Les employés échantillonnés ont été affectés au hasard à la collecte des données par courrier électronique ou par la poste, et des méthodes comparables ont été suivies pour le contact préalable et le suivi des sujets. Le taux de réponse a été un peu plus élevé pour l'enquête par la poste que pour celle par courrier électronique, mais la qualité des données (éléments de données manquants) était la même pour les deux modes d'enquête.

problème, de Leeuw et van der Zouwen (1988) ont procédé à une méta-analyse à grande échelle de 28 études empiriques importantes comportant une comparaison entre études réalisées entre 1952 et 1986 – la plupart aux États-Unis et certaines en Europe – portant sur divers sujets. Les indicateurs de qualité des données utilisés étaient la validité de la réponse (établie d'après des études de validation), l'absence de biais dû au caractère socialement désirable, la réponse à une question, la quantité d'information (pour les questions ouvertes ou les listes de vérification) et la similitude des réponses. Dans l'ensemble, l'analyse montre que, s'il existe des différences de qualité entre les deux modes d'interviews, celles-ci sont définitivement très faibles et que d'autres aspects, comme les coûts et la commodité, devraient être pris en considération pour décider de recourir à l'interview téléphonique pour le travail d'enquête. Des conclusions similaires sont tirées pour le Royaume-Uni par Sykes et Collins (1988) d'après quatre études comparatives, pour le Danemark par Körmendi (1988) pour les données sur le revenu d'après une étude de validation fondée sur des données admnistratives et pour le Canada par Caron et Lavallée (1998), l'Enquête financière sur les fermes.

D'autres études récente sur les effets du mode d'interview se concentrent sur des problèmes et des sujets particuliers, mais aboutissent aux mêmes conclusions. Ainsi, Herzog et Rodgers (1988) présentent les résultats d'une comparaison des modes d'interview dans le cas d'une étude auprès de personnes âgées et n'observent que de légères différences. Foley et Brook (1990) présentent des résultats similaires pour une enquête sur les derniers jours de la vie. Lors d'une étude sur le sujet délicat de la consommation de drogues, Aquilino et Lo Sciuto (1990) obtiennent des résultats presque identiques pour les personnes de race blanche, mais observent certains écarts significatifs pour les personnes de race noire, même après avoir tenu compte de l'effet de variables éventuellement liées au sous-dénombrement téléphonique. Les résultats d'une enquête téléphonique sur la consommation de drogues présentée par Johnson, Fendrich, Shaligram et Garey (1997), qui étayent un modèle des effets d'interview axé sur la distance sociale pourraient expliquer cette situation.

Il y a peu de doutes que l'effet de l'intervieweur sur la qualité des données soit important, que l'enquête soit réalisée sur place ou par téléphone. L'utilisation d'installations centralisées d'interview téléphonique permet de mieux contrôler et surveiller les effets d'intervieweur que dans le cas de l'interview sur le terrain. À cet égard, certaines questions sont traitées par Stokes et Yeh (1988) qui proposent un modèle bayésien des effets d'intervieweur, ainsi que des méthodes d'estimation des paramètres du modèle. Pannekoek (1988) propose un modèle bêta-binomial de la composante de la variance due à l'intervieweur et des méthodes d'estimation des paramètres du modèle.

Un moyen efficace de réduire les erreurs de réponses dans le cas des enquêtes avec interview sur place consiste à utiliser des dossiers fournis par les répondants pour vérifier et rappeler les renseignements sur le revenu, les assurances, les événements liés à la santé, etc. Manifestement, l'extension de cette méthode à l'interview téléphonique pose certains problèmes, puisque l'intervieweur ne peut consulter lui-même les documents et que même le fait de demander aux répondants d'aller chercher ces documents peut causer un arrêt plus fréquemment perturbant dans le cas de l'interview téléphonique que dans celui de l'interview sur place. Cependant, l'utilisation de dossiers par les répondants durant les enquêtes téléphoniques peut contribuer à réduire le biais de réponse. Battaglia, Shapiro et Zell (1996) décrivent une expérience consistant à demander aux répondants d'utiliser les dossiers de vaccination durant l'un des cycles de la US National Immunization Survey et à comparer les renseignements recueillis à ceux des dossiers des vacinateurs. Bien que quelque 47 % de répondants se soient effectivement servi de leur dossier de vaccination, un biais important de sous-dénombrement persiste, peut-être parce que les rapports de vaccination ne sont pas toujours à jour. Des effets comparables ont été constatés pour les enquêtes réalisées sur place – consulter Brick, Katton, Nixon, Givens et Ezzati-Rice (2000).

4. PROGRÈS TECHNOLOGIQUES COURANTS ET FUTURS

Conjugués à la couverture téléphonique presque complète, le progrès technologique très intense et la diversité des moyens de communication ne cessent d'offrir de nouvelles occasions d'utiliser des moyens neufs de communication pour réaliser les enquêtes. Par ailleurs, certains de ces progrès pourraient compléter l'application des téléenquêtes selon la méthodologie classique appliquée aujourd'hui. Ainsi, la complexité croissante des dispositifs et des algorithmes de filtrage (comme l'évolution des simples répondants et dispositifs d'identification des appelants mentionnés à la section 3.3) permettra peut-être aux répondants d'éviter plus facilement de coopérer. Nous allons maintenant examiner les applications courantes, ainsi que les développements et les applications futurs prévus, et commentier les problèmes méthodologiques que pose leur utilisation.

4.1 Courrier électronique et enquêtes par Internet

Le nombre de ménages qui ont accès à Internet a augmenté très rapidement ces dernières années. Par exemple, aux États-Unis, la proportion de ménages raccordés à Internet est passée de 26 % en décembre 1998 à 42 % en août 2000 (NTIA 2000). Dans d'autres pays, la proportion est un peu plus faible, soit 28 % au Royaume-Uni (en août 2000 – OFTEL 2000), 25 % au

un ménage peut causer une sélection avec probabilités inégales. Le cas échéant, si l'on recueille des renseignements sur le nombre de lignes téléphoniques auxquelles le ménage est rattaché, la correction est simple. Par conséquent, une pondération est nécessaire pour tenir compte des UPF pour lesquelles le nombre de numéros faisant partie du champ d'observation est inférieur à la taille requise de l'échantillon de grappes. Le fait qu'il soit souvent difficile de déterminer si un numéro de téléphone pour lequel on n'obtient aucune réponse après des tentatives répétées d'appel est effectivement un cas de non-réponse, d'un numéro qui fait partie du champ d'observation ou, en fait, un numéro hors du champ d'observation pose un problème supplémentaire. La non-réponse, le sous-dénombrement inhérent aux ménages qui ne possèdent pas le téléphone et la nécessité manifeste d'utiliser une certaine forme d'estimateur de la multiplicité dans le cas des plans d'échantillonnage à bases de sondage multiples, en s'appuyant sur les données des bases de sondage dans lesquelles l'unité est représentée, sont d'autres problèmes qu'il faut résoudre.

Ces problèmes sont traités, dans le cas de l'échantillonnage national par CA réalisé par le US National Center for Health Statistics, dans une série d'articles publiés par Thornberry et Massey (1978), Botman, Massey et Shimizu (1982), ainsi que Massey et Botman (1988). Ces articles décrivent les rajustements de la pondération effectués pour la US National Health Interview Survey (NHIS) réalisée par CA et pour une enquête sur l'usage du tabac afin de tenir compte de la multiplicité des lignes ou des numéros de téléphone par ménage, de la couverture téléphonique et de la non-réponse. Ces rajustements se fondent sur des données externes pour la race et la région géographique, et sur des données d'enquête pour la non-réponse et les lignes téléphoniques multiples. Plusieurs méthodes de rajustement et de pondération sont comparées et évaluées. Chapman et Roman (1985) comparent la substitution à la correction pour la non-réponse lors d'une étude de faisabilité ayant trait à la NHIS par CA et constatent que le biais et la variance sont comparables. Drew et Groves (1989) comparent diverses méthodes de rajustement pour la non-réponse unitaire fondées sur l'utilisation de données administratives externes au moyen d'un modèle explicite de prédiction de la réponse et sur les probabilités de réponse estimées d'après les données concernant les appels de suivi. Casady et Sirken (1980) proposent un estimateur de multiplicité pour un plan d'échantillonnage à bases de sondage multiples applicable aux données de la US National Health Interview Survey. Brick (1990) compare l'estimateur de multiplicité à l'estimateur à bases multiples classique dans le cas d'une enquête par CA sur la scolarité.

Grosec, Juddkins et Mosher (1991) décrivent des rajustements, fondés sur la modélisation de la propension à la non-réponse, dans le cas d'un suivi téléphonique après une interview sur place de la US National Survey of Family Growth. Bull, Pederson et Ashley (1988)

proposent une correction fondée sur la propension à répondre selon l'intensité des efforts de suivi et selon la catégorie d'usage du tabac pour une enquête canadienne sur les attitudes à l'égard des dispositions législatives limitant l'usage du tabac.

À la suite d'une comparaison des ménages qui n'ont pas le téléphone aux ménages « en transition » (c'est-à-dire ceux qui ont acquis ou perdu récemment le service téléphonique) faite par Keeter (1995), Brick, Wakseberg et Keeter (1996) proposent d'utiliser les données sur l'interruption du service téléphonique pour faire la correction pour le sous-dénombrement du aux ménages qui n'ont pas le téléphone. Les résultats montrent que ce genre de correction peut réduire l'erreur quadratique moyenne. Hoaglin et Battaglia (1996) comparent une méthode modifiée de stratification avec stratification simple pour corriger les données de la non-couverture dans le cas d'une enquête par CA sur la couverture de la vaccination. La méthode modifiée de stratification a posteriori s'appuie sur des données nationales sur les taux de vaccination pour les enfants qui vivent dans des ménages ayant et n'ayant pas le téléphone, ainsi que sur des données démographiques et socioéconomiques utilisées pour la stratification a posteriori simple, tandis que, dans le cas de la méthode basée sur un modèle, on se fonde sur un modèle logit pour estimer la probabilité de vivre dans un ménage ayant le téléphone. L'étude montre que l'utilisation de la méthode modifiée de stratification a posteriori produit une amélioration, mais que les résultats de la méthode modifiée de stratification a posteriori et de la correction basée sur un modèle diffèrent peu. Frankel, Srinath, Battaglia, Hoaglin, Wright et Smith (1999) appliquent une correction similaire, fondée sur des données de la NHIS et montrent de façon concluante qu'elle réduit considérablement le biais.

3.4 Qualité des données – Erreurs de réponse et effets de mode

La qualité des renseignements recueillis par téléphone a toujours été une question controversée. Comme on l'a mentionné à la section 2, les craintes quant à la qualité view téléphonique ont été apaisées très tôt, en grande partie grâce à certaines évaluations empiriques à grande échelle réalisées durant les années 1960 et 1970. Cependant, certaines données contradictoires provenant de diverses études sur la qualité relative des interviews par téléphone et sur place persistaient néanmoins. Même si l'analyse approfondie des données de grandes enquêtes omnibus réalisées selon les deux modes d'interviews par le University of Michigan Survey Research Center (Groves et Kahn 1979) ont fourni des renseignements importants sur la qualité des données et d'autres questions, les comparaisons entre les modes d'interview et la comparaison à des données externes n'ont pas été concluantes. En vue de résoudre le

d'un répondeur, ces personnes sont au moins aussi susceptibles de participer à l'enquête que les personnes qui ne possèdent pas de répondeur. Il fait également remarquer que tomber sur un répondeur donne la certitude qu'un ménage a été rejoint et que ses membres ne veulent pas manquer les appels importants. Selon une étude conçue par Xu, Bates et Schweitzer (1993) pour étudier l'effet des messages laissés sur les répondeurs, les ménages qui possèdent un répondeur sont plus susceptibles d'être contactés et de participer à l'interview que ceux qui n'en possèdent pas. De surcroît, le fait de laisser un message sur le répondeur augmente de façon significative le taux de réponse et réduit de façon significative le taux de refus.

L'intervieweur et de la ville de résidence. Selon un essai randomisé réalisé par Koepsell, McGuire, Longstrech, Nelson et van Belle (1996), laisser un message sur le répondeur produit une augmentation globale de 20 % du taux de réponse. Bien que, lors d'une étude similaire, Tucker et Shukers (1997) n'aient constaté aucun effet significatif, les résultats globaux d'un éventail d'études indiquent que l'usage accru des répondeurs a un effet favorable sur le taux de réponses aux enquêtes, probablement parce qu'il donne la possibilité de laisser un message positif, donc de permettre aux personnes appelées de filtrer les appels des télévendeurs.

Tucker et O'Neill (1996) estiment que la proportion de ménages américains possédant un dispositif d'identification des appelants est passé de 3 % en 1992 à 10 % en 1996. Selon une étude nationale, qui comprend l'analyse du profil des abonnés à l'option d'identification des appelants et des propriétaires de répondeurs, ces auteurs concluent que ces dispositifs technologiques ne constituent pas encore des obstacles importants à la recherche par sondage téléphonique, puisque leurs propriétaires ont tendance à utiliser les dispositifs de filtrage principalement pour rejeter les appels indésirables provenant de personnes dont ils connaissent le numéro plutôt que les appels provenant de numéros inconnus. Cependant, ils soulignent que la possibilité de filtrer les appels augmentera probablement le taux de réponse par répondeur aux appels de suivis répétés.

3.3.3 Pondération et rajustement des données

Une attention particulière doit souvent être accordée à la pondération et au rajustement des données recueillies par enquête téléphonique. Bien que les plans d'échantillonnage se fondent habituellement sur la sélection avec probabilités égales, en pratique ces conditions ne sont pas toujours atteintes. Par exemple, en théorie, les plans d'échantillonnage par CA sont autopondérés, mais, en réalité, la multiplicité des lignes téléphoniques (numéros) que possède

indique que le taux de réponse est nettement plus faible pour les enquêtes téléphoniques que pour les enquêtes sur place si l'on applique des modèles à pente constante. Cependant, si l'on choisit des modèles à pente aléatoire, l'écart n'est plus significatif.

Certains chercheurs ont étudié l'effet des variables opérationnelles d'enquête sur la non-réponse pour essayer de réduire cette dernière dans le cas des enquêtes téléphoniques. Ainsi, lors d'une expérience en prévision de la US National Crime Survey, Sebold (1988) constate que le fait de doubler la période d'enquête (pour passer de deux à quatre semaines) augmente le taux de réponse de trois points de pourcentage. Brick et Collins (1997) étudient l'effet de l'envoi préalable de lettres et des questions de filtrage sur le taux de réponse à la US National Household Education Survey. Selon eux, l'adoption d'une méthode de rejet à la sélection augmente considérablement le taux de réponse, mais l'envoi d'une lettre préalable ne rehausse pas l'effet de la présélection. La durée de l'interview (Collins et coll. 1988) et les caractéristiques vocales de l'intervieweur (Oksenberg et Cannel 1988) sont d'autres variables qui influent sur le taux de réponse. L'effet de la méthode de sélection de personnes dans le ménage sur la non-réponse (en particulier la nécessité d'obtenir la liste des membres du ménage) a déjà été mentionnée à la section 3.2.2.

Enfin, ces dernières années, on a assisté à une augmentation significative de l'utilisation de répondeurs et de dispositifs d'identification des appelants en vue de filtrer les appels non désirés, situation qui augmente la probabilité de non-réponse. Par exemple, en France, la proportion de ménages équipés d'un répondeur est passée de 21 % en 1995 à 40 % en 1999 (Rouquette 2000) ; l'augmentation est la même en Allemagne (République fédérale d'Allemagne 1999), tandis qu'aux États-Unis, la proportion est passée d'environ 25 % en 1988 (Tucker et Feinberg 1991) à plus de 73 % en 1997 (*Decision Analyst* 1997). Cependant, d'après une enquête téléphonique nationale, Tucker et Feinberg (1991) concluent que, comparativement à d'autres groupes de non-réponse initiale (par exemple, « pas de réponse » ou « signal occupé », les ménages dotés d'un répondeur sont plus susceptibles de répondre et moins susceptibles de refuser de participer à l'enquête, ce qui produit un taux de réponse pour ces ménages qui n'est définitivement pas plus faible que celui observé pour les autres formes de non-réponse. En fait, selon une étude réalisée par Oldendick et Link (1994), il semble que le taux d'utilisation d'un répondeur pour filtrer les appels des enquêteurs est de l'ordre de 2 % à 3 % seulement. Cependant, les personnes qui filtrent effectivement les appels ont tendance à appartenir aux catégories supérieures de revenu, à vivre en région urbaine et à avoir atteint un niveau élevé de scolarité. Pareillement, Piazza (1993) constate, après examen de la foule de données provenant de la California Disability Survey, enquête téléphonique pour laquelle le nombre de rappels est élevé, que, s'il est plus difficile au départ d'entrer en contact avec les propriétaires

« pages jaunes » permet d'éliminer a priori un grand nombre de numéros commerciaux. Cette méthode et d'autres réduisent le coût du filtrage et l'ambiguïté ayant trait aux appels qui restent sans réponse.

Les progrès technologiques, tels que le « renvoi automatique d'appels » et l'identification de l'appelant, favorisent la non-réponse. En outre, il est plus facile de refuser ou de mettre fin à une interview téléphonique qu'à une interview en personne. Groves et Lyberg (1988a) ont étudié en profondeur ce problème de non-réponse et d'autres que posent l'interview téléphonique « à froid » et les moyens mis en œuvre aux États-Unis pour les résoudre. Plus précisément, comme l'ont fait CASRO (1982) et White (1983), ils recommandent que la définition du taux de non-réponse inclue, au dénominateur, une estimation du nombre de numéros restés sans réponse qui sont des numéros en service, en plus du nombre d'interviews complètes et incomplètes, du nombre de numéros qui ont produit un refus et du nombre d'autres unités non interviewées. L'estimation du nombre de numéros admissibles qui n'ont produit aucune réponse est exprimée sous forme de proportion par rapport au nombre de numéros admissibles qui ont produit une réponse. Cependant, cette estimation pourrait être biaisée. Par exemple, l'usage intensif de répondants par les entreprises sous-entend que presque toutes les entreprises répondront et pourront être identifiées comme étant des entreprises. En outre, comme le fait remarquer Massey (1995), dans le cas du filtrage, ou sélection par tri, il faut modifier cette mesure en définissant, pour le filtrage des ménages, un taux de réponse qui correspond à la proportion estimative de ménages admissibles identifiés comme tels lors du tri de sélection, plutôt que sous forme de proportion de l'ensemble de ménages filtrés pour déterminer leur admissibilité. Cunningham, Brick et Meader (2000) présentent plusieurs mesures détaillées du taux de réponse et du taux d'admissibilité pour chaque étape d'une enquête avec filtrage, ainsi que du taux de réponse global, lors de la description de la méthodologie de la National Survey of America's Families.

En général, les taux de non-réponse sont plus élevés dans le cas des enquêtes téléphoniques que dans celui des enquêtes sur place, pour les raisons susmentionnées – consulter Hochstim (1967), Groves et Kahn (1979), Fitt (1979), Groves et Lyberg (1988) pour la situation aux États-Unis, Wilson, Blackshaw et Norris (1988); et Collins, Sykes, Wilson et Blackshaw (1988) pour la situation au Royaume-Uni, ainsi que Drew, Choudry et Hunter (1988) en ce qui concerne les enquêtes gouvernementales au Canada. Ces derniers auteurs procèdent aussi à une comparaison des interviews téléphoniques « à froid » et « à chaud » qui ne révèle que de faibles écarts entre les taux de non-réponse. Plus récemment, une analyse de la situation pour 39 enquêtes téléphoniques américaines réalisées durant les années 1990 (Massey, O'Connor et Krotki 1997) a montré une légère réduction supplémentaire du taux de réponse, la moyenne étant de 62 % et la fourchette variant

de 42 % à 79 % (toutefois, il semble qu'au Canada, les taux de réponse n'aient pas diminué ces dernières années). L'utilisation croissante de dispositifs technologiques (répondeurs, renvoi automatique des appels, lignes téléphoniques multifonctionnelles) et la fréquence croissante de la sollicitation par téléphone, déjà reconnue par Biel (1967) comme pouvant poser des problèmes pour les enquêtes téléphoniques, comptent parmi les facteurs auxquels on peut imputer cette augmentation de la non-réponse. L'American Statistical Association (1999) estime que la diminution du taux de participation aux enquêtes causée par la quasi saturation provoquée par les appels des entreprises de telemarketing est un problème grave auquel les spécialistes des études par sondage ne se sont pas attaqués pleinement. L'Association conclut que, si la tendance ne se renverse pas, les enquêtes téléphoniques telles que nous les connaissons disparaîtront dans les cinq ans à venir.

Kalton (2000) est du même avis.

Comme cela est le cas pour la non-couverture téléphonique, le biais dû à la non-réponse introduit dans les estimations d'enquête est encore accentué par la corrélation entre la non-réponse et nombre de caractéristiques socio-économiques. Groves et Lyberg (1988a) se sont fondés sur l'examen de travaux antérieurs pour cerner les principaux corrélats de la non-réponse téléphonique. Il s'agit de l'âge (le taux de refus est plus élevé chez les personnes âgées – consulter aussi Collins et coll. 1988) et le niveau de scolarité (le taux de réponse est plus élevé pour les groupes dont le niveau de scolarité est faible – consulter, par exemple, Cannel, Groves, Magilav, Mathiowetz, Miller et Thornberry 1987). Par contre, il existe des preuves que l'écart entre les taux de non-réponse observés pour les régions urbaines et rurales est plus faible dans le cas des enquêtes téléphoniques que dans celui des enquêtes sur place (Groves et Kahn 1979). Des articles plus récents sur les effets de la non-réponse se concentrent sur des problèmes particuliers. Ainsi, Diehr, Collins, Sykes, Wilson et Blackshaw (1992) étudient le lien entre le taux de réponse et d'autres variables sommatoires au niveau du préfixe et de la personne et notent que la non-réponse est corrigée à l'âge, à la race, à la taille de la famille et au type de cette dernière. Dans une étude de l'effet des appels de suivi sur la qualité des estimations d'enquête, Merkle, Bauman et Lavrakas (1993) montrent que l'âge et la situation d'emploi sont les principales variables corrigées au nombre de rappels. Kalsbeek et Durham (1994) étudient l'effet de la non-réponse dans le cadre d'une enquête téléphonique de suivi sur l'alaitement maternel auprès de femmes à faible revenu et constatent que la non-réponse est corrigée principalement à l'âge et au degré d'urbanisation. Enfin, Hox, DeLeeuw et Kreft (1991) recourent à la modélisation multilevel pour réaliser une méta-analyse de grande portée des rapports sur les comparaisons de la non-réponse selon le mode d'enquête. Leur étude, fondée sur l'analyse par modélisation multilevel de 45 études (dont 35 comportant une composante téléphonique),

déterminer le nombre total de numéros de téléphone (services mobile et fixe) que possède un ménage (information nécessaire pour la pondération) peut être une tâche fastidieuse. À la section 4, nous considérons certains moyens éventuels d'aborder cette question et d'autres problèmes que pose le passage à la téléphonie mobile.

Le sous-dénombrement des personnes dans les ménages couverts a trait principalement à la méthode de sélection des personnes dans le ménage (voir la section 3.2.2) et au sous-dénombrement dû au fait que l'on n'obtient pas toujours la liste complète des personnes qui composent le second facteur en comparant les données sur les particuliers provenant d'une enquête par CA à celles provenant de la US Current Population Survey et du Recensement de la population. Selon ces auteurs, bien que la taille moyenne des ménages soit comparable, les résultats de l'enquête par CA sont biaisés en faveur des ménages de deux personnes et en défaveur des ménages d'une seule personne. Les écarts pourraient être dus en partie à l'application de règles de résidence différentes, mais les résultats n'indiquent pas que les personnes sont sous-dénombrées dans le cas de l'enquête par CA. Ils décrivent aussi une expérience qui a consisté à poser des questions plus détaillées sur la composition du ménage et ne constataient presque aucune amélioration de l'exactitude de la déclaration. Lors d'une enquête sur l'usage du tabac, Bercini et Massey (1979) ont testé les effets de l'utilisation de noms dans la liste de composition du ménage et de la position de la question sur la composition du ménage (avant ou après la première interview). Selon eux, l'utilisation des noms et la position de la question sur la composition du ménage ont toutes deux un effet sur la réponse et l'obtention de la composition du ménage après l'interview est la méthode qui donne les meilleurs résultats.

3.3.2 Non-réponse

La non-réponse et le biais qui y est associé est un problème fondamental de toutes les études par sondage, mais l'interview téléphonique pose des problèmes particuliers de non-réponse. L'un des problèmes principaux tient à l'ambiguïté des résultats de nombreux essais de composition – par exemple, ligne continuellement occupée ou pas de réponse, numéro raccorde à un télécopieur, un modem d'ordinateur ou un répondeur. Récemment, des dispositifs automatisés de filtrage ont été mis au point pour repérer les numéros de téléphone raccordés à des enregistrements indiquant s'ils sont ou non en service (Casady et Lepkowski 1999). Ainsi, du matériel et des logiciels personnalisés ont été développés pour déceler les enregistrements « à trois tonalités » qui indiquent que le numéro « n'est pas en service » de sorte que, s'ils sont composés, ces numéros puissent être éliminés de l'échantillon. L'appariement des listes de numéros aux fichiers des

Starer 1996; Fox et Riley 1996; NTIA 2000). Selon Anderson, Nelson et Wilson (1998), qui se sont fondés sur les données de la National Health Interview Survey, et selon Ford (1998) qui s'est fondé sur les données de la National Health and Nutrition Examination Survey, les caractéristiques liées à la santé des personnes qui vivent dans un ménage ayant le téléphone diffèrent quelque peu de celles des personnes qui vivent dans un ménage n'ayant pas le téléphone. Cependant, les deux études mènent à la conclusion que les effets de la couverture téléphonique sont faibles.

Par contre, le problème principal que pourrait poser la couverture téléphonique dans un avenir proche a trait au lancement et à la prolifération rapide de la téléphonie mobile. À la fin des années 1990, la proportion de ménages ayant accès à au moins un téléphone mobile atteignait 76 % en Finlande, 59 % au Danemark, 35 % en Italie (Rouquette 2000) et 52 % en Israël (Bureau central de la statistique 2000). La situation ne poserait aucun problème si ces téléphones mobiles étaient simplement complètement de service fixe existant. Toutefois, on possède déjà des preuves convaincantes d'une tendance, dans plusieurs pays, à considérer le téléphone mobile comme un remplacement, plutôt qu'un complément, du service téléphonique fixe. Selon Kuusela et Viikari (1999), en Finlande, 20 % des ménages possèdent maintenant un ou plusieurs téléphones mobiles, mais ne possèdent aucun téléphone fixe et dans un an, le nombre de téléphones mobiles surpassera le nombre de lignes téléphoniques fixes. Comparativement, les chiffres sont de 3 % pour le Royaume-Uni (OFTEL 2000) et de 2,9 % pour Israël (Bureau central de la statistique 2000). Autrement dit, la couverture des services téléphoniques fixes a diminué pour s'établir à 77 % en Allemagne, on estime que la proportion de ménages abonnés à un service téléphonique fixe diminuera pour s'établir à 92 % d'ici à 2004 (Gabler et Haeder 2000). De surcroît, les caractéristiques des personnes qui possèdent uniquement un téléphone mobile sont assez différentes de celles qui s'abonnent à un service téléphonique fixe. D'après Kuusela et Viikari (1999), en Finlande, les premières sont généralement jeunes, vivent souvent seules, en appartement, en région urbaine. Il convient de souligner que le passage de la téléphonie fixe à la téléphonie mobile ne semble pas se produire dans les mêmes proportions en Amérique du Nord, en raison de stratégies différentes d'établissement des prix.

En théorie, l'échantillonnage par CA pourrait être étendu à la téléphonie mobile. Toutefois, en pratique, l'exercice pourrait être assez difficile, parce que le téléphone mobile est, de par sa nature, un appareil personnel plutôt qu'un appareil du ménage. Échantillonner des personnes dans un ménage en communiquant avec l'un des membres par téléphone mobile est virtuellement impossible. Interviewer par téléphone mobile des personnes qui peuvent se trouver n'importe où est également une tâche fort difficile. Même

personne prédéterminée (par exemple, « l'homme le plus âgé ». De nouveau, cette méthode ne permet pas d'assurer que la probabilité de sélection soit positive pour tous les

membres du ménage.

Les méthodes susmentionnées ont fait l'objet de plusieurs comparaisons empiriques. Czaja, Blair et Sebestik (1982) ne constataient aucune différence significative entre les taux de réponse ni les profils démographiques des deux versions de la méthode de Trol Dahl-Carter et de la méthode de Kish. Hagan et Meier (1983) comparèrent leur méthode, décrite plus haut, à celle de Trol Dahl-Carter et constatèrent que la méthode qu'ils proposent produit un taux de refus nettement plus faible, sans différence significative entre les profils démographiques. Salmon et Nichols (1983) comparèrent quatre méthodes de sélection des répondants dans un ménage – méthodes de Trol Dahl-Carter, d'alternance homme/femme, du prochain universaire et de sélection nulle, dans le cadre d'une enquête téléphonique auprès d'un petit échantillon. Ils conclurent que la méthode du prochain universaire est assez efficace pour sélectionner un échantillon représentatif des membres du ménage. Oldendick, Bishop, Sorenson et Tuchfarber (1988) ne constataient aucune différence significative entre la méthode de Kish et celle du dernier universaire. Lors d'une étude fondée sur la méthode du dernier universaire, Romuald et Haggard (1994) observèrent que le taux d'autosélection des personnes bien informées en vue de participer à l'enquête est plus élevé que prévu. Ils étudiaient l'effet qu'ont les indices utilisés pour raviver la mémoire sur l'autosélection du répondant et conclurent que cet effet n'est pas significatif. Lavarakas, Bauman et Merkle (1993) évaluaient l'effet de l'utilisation de la méthode du dernier universaire sur la couverture à l'intérieur d'une unité dans le cas d'une enquête nationale et présentaient des données qui donnaient à penser que la méthode aboutit à une sélection incorrecte dans de nombreux cas. Forsman (1993) passe en revue les expériences d'échantillonnage dans les ménages réalisées par 18 sociétés spécialisées dans les sondages d'opinion et décrivent un test permettant de comparer les méthodes de Kish, du prochain/dernier universaire et de Trol Dahl-Carter. Ils conclurent que cette dernière méthode est un peu meilleure que la méthode de Kish et que toutes deux sont supérieures à la méthode des annuaires. Pareillement, Binson, Cancho et Catania (2000) décrivent, pour une enquête téléphonique nationale, une comparaison triple entre les méthodes de Kish, du prochain universaire et du dernier universaire et notent des écarts significatifs entre les taux d'abandon enregistrés pour les trois méthodes aux premiers stades du processus de filtrage. La méthode de Kish produit le taux d'abandon le plus élevé et celle du « prochain universaire », le taux le plus faible. Ils supposent que les intervieweurs, plutôt que les répondants, sont la source principale du taux élevé de refus dans le cas de la méthode de Kish, puisque cette dernière nécessite la liste complète des membres du ménage.

3.3 Couverture et non-réponse

3.3.1 Couverture téléphonique

Jusqu'à récemment, le problème de la non-couverture téléphonique était un inconvénient important des enquêtes par téléphone. Même aux États-Unis, le sous-dénombrement des personnes (dans les ménages ne possédant pas le téléphone) était encore de 7,2 % à la fin de 1986 (Thomberry et Massey 1988). Au milieu des années 1980, le sous-dénombrement téléphonique des ménages était inférieur à 10 % dans la plupart des pays occidentaux, le taux de couverture le plus élevé (99 %) étant celui observé en Suède. Néanmoins, le taux de sous-dénombrement téléphonique demeurerait élevé dans certains pays, comme le Royaume-Uni (25 %), l'Italie (29 %), l'Irlande (50 %) et Israël (30 %) (Trewin et Lee, 1988). À la fin du siècle, la situation avait changé spectaculairement, puisque, dans la plupart des pays occidentaux, la couverture téléphonique atteignait virtuellement la saturation. Elle était de 94,4 % aux États-Unis en 1999 (NTIA 2000), de 96,6 % en Australie en 1996 (St. Clair et Muir 1997), de 97,0 % au Royaume-Uni (OFTEL 1999), de 97,3 % en Israël (Bureau central de la statistique 2000) de 97,9 % en Finlande (Kauusela et Vikki 1999), de 98,2 % au Canada (Statistique Canada 1999) et de 99 % en Allemagne (République fédérale d'Allemagne 1999).

De toute évidence, le principal problème que pose le sous-dénombrement téléphonique tient davantage au sous-dénombrement différentiel qu'au sous-dénombrement global, et au fait que le sous-dénombrement téléphonique est fortement corrélé au large éventail de variables démographiques, économiques et de santé. Cette situation a été confirmée par un grand nombre d'études empiriques menées aux États-Unis et ailleurs – à cet égard, consulter par exemple Groves et Kahn (1979), Collins (1983, 1999), Thomberry et Massey (1983, 1988), Trewin et Lee (1988), ainsi que Botman et Allen (1990). L'augmentation rapide de la couverture téléphonique globale au cours de la dernière décennie n'a pas modifié radicalement cette situation. Ainsi, en Finlande, où le sous-dénombrement téléphonique global était de 2,1 % en 1999, le sous-dénombrement des ménages à faible revenu (moins de 675 Euros par mois) était de 11,3 % (contre 0 % pour les ménages à revenu élevé) et celui des ménages vivant dans un logement loué, de 4,9 % (Kauusela et Vikki 1999). En Israël, le sous-dénombrement téléphonique était de 17,9 % pour le décile inférieur de revenu comparativement à 0,8 % pour les deux déciles supérieurs et de 24,9 % pour les ménages composés d'un seul adulte et d'au moins trois enfants comparativement à 2,4 % pour les ménages sans enfant comptant au moins trois adultes (Bureau central de la statistique 2000). Parallèlement, aux États-Unis, les variations géographiques sont importantes et le sous-dénombrement téléphonique est corrélé au problème de logement, à la race, au niveau de scolarité, au revenu et à la mobilité (Shapiro, Battaglia, Hoaglin, Buckley et Massey 1996; Giesbrecht, Kuip et

considérablement les coûts de dénombrement comparativement à l'interview sur place.

3.2.2 Échantillonnage de personnes dans les ménages

Presque toutes les enquêtes-ménages comprennent des questions sur des personnes qui vivent dans le ménage. Dans certains cas, tous les membres du ménage sont inclus dans l'échantillon, mais souvent, pour diverses raisons, on sélectionne un ou plusieurs membres du ménage auxquels on demande de répondre personnellement à certaines questions. La méthode classique de Kish (Kish 1949), utilisée principalement lors des enquêtes par interview sur place, pose des problèmes particuliers dans le cas des enquêtes téléphoniques, car elle oblige à obtenir la liste complète des membres du ménage par téléphone. Ces renseignements sont plus difficiles à obtenir par téléphone que lors d'une interview sur place durant laquelle certaines personnes peuvent être présentes. Il convient toutefois de souligner que, dans de nombreux cas, il faut de toute façon recueillir les renseignements sur la composition du ménage. En outre, la manipulation des règles de sélection par l'intervieweur (par exemple, pour obtenir un taux de réponse élevé) souponne depuis longtemps dans le cas des interviews sur place, est presque impossible dans le cas des enquêtes par ITAO (pour lesquelles l'intervieweur ne peut visualiser la sélection).

Troldahl et Carter (1964) proposent une méthode qui n'oblige à déterminer que le nombre de personnes de chaque sexe dans le ménage. Puis, des règles probabilistes (par exemple, l'homme le plus âgé) sont appliquées pour sélectionner la personne qui répondra aux questions, de sorte que l'on connaît les probabilités de sélection de chaque personne. Cependant, la probabilité de sélection n'est pas positive pour tous les membres du ménage (par exemple, dans les ménages comptant trois hommes, celui d'âge intermédiaire n'est jamais sélectionné). La méthode (appelée « méthode de Troldahl-Carter ») a été modifiée par Bryant (1975) afin de tenir compte du fait que des ménages peuvent compter plus de deux personnes de même sexe. Une variante de la méthode, proposée par Salmon et Nichols (1983) et par O'Rourke et Blair (1983), consiste à sélectionner la personne qui sera la prochaine (ou a été la dernière) à célébrer son anniversaire (méthode du « prochain anniversaire » ou du « dernier anniversaire »), pour s'assurer que la probabilité de sélection soit la même pour tous les membres du ménage, en supposant que la date de l'interview soit aléatoire. Évidemment, cette hypothèse est raisonnable dans le cas des enquêtes réalisées sur une période de 12 mois, mais ne l'est pas pour les enquêtes dont la période d'interview est plus courte. Ce facteur et d'autres pourraient donner lieu à une corrélation entre les probabilités de sélection et les caractéristiques individuelles. Une autre méthode de sélection proposée par Hagan et Meier (1983), ne nécessite aucune donnée préliminaire sur la composition du ménage et consiste à sélectionner une

Le coût relativement faible de l'interview téléphonique fait de celle-ci un moyen de premier choix de filtrer les grands échantillons en vue de repérer de petites populations particulières. Ainsi, Sudman (1978) discute des conditions dans lesquelles l'utilisation d'un échantillon téléphonique pour la sélection par tiré d'un sous-groupe dont les membres seront, en bout de ligne, interviewés sur place est plus efficace que le filtrage sur place. L'analyse des fonctions de coût montre que le filtrage téléphonique est efficace, sauf si l'homogénéité intragroupe est faible, la densité d'interview est faible et/ou les coûts de dépistage et de filtrage sont faibles comparativement aux coûts de l'interview. Blair et Czaja (1982) proposent une modification de la méthode de Mitofsky-Waksberg pour repérer les populations spéciales géographiquement regroupées et décrivent une application à la population noire. Toutefois, comme le fait remarquer Waksberg (1983), quand les grappes sont épuisées, cette méthode nécessite une reproduction qui pourrait réduire son efficacité. Autrement dit, la méthode pourrait être efficace pour la population noire, mais ne pas l'être nécessairement pour d'autres minorités. Un autre plan de sondage téléphonique applicable à la population noire des États-Unis est proposée par Inglis, Groves et Heeringa (1987). Mohadjer (1988) propose un plan d'échantillonnage par CA avec stratification des zones de préfixe. Hartge, Brinton, Rosenthal, Cahill, Hoover et Waksberg (1984) décrivent la sélection de ménages par la méthode de Mitofsky-Waksberg conjuguée à l'échantillonnage stratifié de personnes dans les ménages pour sélectionner un groupe témoin représentatif de la population dans le cas de quatre études épidémiologiques. Une simulation de l'échantillonnage aléatoire simple en vue de déterminer l'efficacité de la méthode porte Permeget, Myers, Klag et Whelton (1993), à conclure que celle-ci est efficace.

Les enquêtes aréolaires sont un autre exemple de population spéciale que l'enquête téléphonique permet de traiter efficacement. Même si, en général, les centraux téléphoniques ne coïncident pas avec des régions géographiques exactes, le degré de correspondance est important et, si l'on procède à une sélection par tiré des unités comprises dans la zone définie, l'interview téléphonique permet de réduire considérablement les coûts. Par exemple, Banks et Hagan (1984) décrivent la réduction du filtrage que doivent réaliser les intervieweurs grâce à la combinaison de l'échantillonnage à partir d'une liste et de l'échantillonnage par CA lors d'une enquête visant à évaluer l'efficacité des programmes de santé dans des zones de services particulières. Particulièrement, Campbell et Palli (1988) testent une combinaison d'échantillonnage par liste et de composition totale (CT) en se servant comme base de sondage de tous les numéros couverts par les circonscriptions téléphoniques correspondant à une région de recensement donnée et constatent que cette méthode réduit

3.2.1 Échantillonnage de populations spéciales

3.2 Autres problèmes d'échantillonnage

études, dont celles de Landon et Banks (1977) et de Mullet (1982), montrent qu'en fait, cette méthode n'est pas dépourvue de biais et que son efficacité est faible.

Forsman et Danielsson (1997) proposent une méthode axée sur un modèle pour l'échantillonnage par la méthode du «nombre en plus», fondée sur l'hypothèse qu'un préfixe comprend un mélange aléatoire de numéros publiés et non publiés. Le modèle, qui a été testé empiriquement, fournit des estimations non biaisées. Ghosh (1984) a proposé une méthode améliorée qui consiste à continuer d'ajouter une unité au dernier numéro de téléphone composé aussi longtemps qu'on n'a pas rejoint un ménage et d'arrêter aussitôt que l'on en rejoint un. Quoiqu'il persiste un biais, celui-ci est plus faible que celui que produit la simple méthode de l'«unité en plus». D'autres méthodes assistées par liste avec composantes CA sont examinées par Potter, McNeill, Williams et Waitman (1991) qui stratifient les préfixes d'après les dénombrements de numéros de téléphone publiés, tout en s'assurant que soient inclus des blocs ne contenant aucun numéro publié.

Brick, Waksberg, Kuip et Starer (1995) proposent une méthode assistée par liste qui permet de surmonter le problème gênant de la nature séquentielle de l'échantillonnage de deuxième phase du scénario de Mitofsky-Waksberg. La méthode consiste à répartir le fichier des préfixes de central (banques 100) en deux strates. La première comprend tous les préfixes de central qui comptent au moins un numéro de téléphone résidentiel publié et la deuxième, ceux qui n'en contiennent aucun. Limiter l'échantillonnage à la première strate réduit spectaculairement la proportion de numéros non résidentiels qu'il faut composer, mais cause un biais de couverture. Ils étudient le biais et concluent que ces méthodes d'échantillonnage tronquées sont efficaces et présentent des avantages opérationnels, tandis que le biais de couverture résultant (environ 4 %) n'est pas très important. La méthode a été utilisée à grande échelle pour remplacer la méthode classique de Mitofsky-Waksberg. Statistique Canada l'appliqua depuis 1991 à l'enquête sociale générale pour la sélection de l'échantillon complet, avec échantillonnage aléatoire simple dans les banques de numéros qui contiennent au moins un numéro résidentiel (Nortis et Paton 1991). La modification de ce plan de sondage inclut la stratification complète des banques de numéros d'après les renseignements fournis par des listes, l'utilisation de la CA simple pour les strates contenant une faible proportion de banques pour lesquelles aucun numéro résidentiel n'est publié et l'application de la méthode de Mitofsky-Waksberg aux strates restantes. Casady et Lepkowski (1993) commentent ce plan de sondage à d'autres plans stratifiés en servant d'un modèle de coûts. Selon leur étude, pour des rapports de coûts faibles (de sélections productives aux sélections improductives), les plans de sondage par CA à deux et à trois strates sont aussi efficaces que la méthode de Mitofsky-Waksberg et, pour les rapports de coûts élevés, ils sont supérieurs.

3.1.5 Plans à plusieurs bases de sondage

En vue d'éliminer certains biais de sous-dénombrement inhérents aux enquêtes téléphoniques dont la base de sondage est un annuaire ou des numéros de téléphone, on a accordé de plus en plus d'attention aux enquêtes à base de sondage double et à mode mixte combinant l'interview par téléphone et l'interview sur place. Ces enquêtes allient l'échantillonnage classique pour les interviews sur place à l'échantillonnage par CA ou d'après un annuaire pour les interviews téléphoniques. Bieemer (1983) a étudié la composition optimale pour ce genre de plans de sondage grâce à une étude en simulation, et McCarthy et Bateman (1988) proposent de recourir à la programmation mathématique pour atteindre une répartition optimale des unités d'échantillonnage pour un plan à base de sondage double qui permet l'analyse a posteriori des effets des variations du plan de sondage et des paramètres de coût sur l'optimisation. Choudhry (1989) propose une optimisation variable selon le coût pour estimer les proportions et Brick (1990) propose pour cela de recourir à l'échantillonnage multiple. Dans une série d'articles, Groves et Lepkowski (1985, 1986), Lepkowski et Groves (1984, 1986b) et Traugott, Groves et Lepkowski (1987) mettent au point des modèles d'erreur pour ces plans d'enquête à base de sondage double. Ils présentent aussi les résultats d'expériences visant à comparer les taux de réponse et les biais éventuels pour l'échantillonnage par CA ou d'après une liste et pour plusieurs méthodes d'interview. Les résultats ont été appliqués aux États-Unis à la grande enquête National Crime Survey.

Whitmore, Mason et Hartwell (1985) décrivent l'application de méthodes à base de sondage et mode doubles lors d'une étude de l'exposition personnelle au monoxyde de carbone dans deux régions métropolitaines parrainée par la US Environment Protection Agency et d'une étude à l'échelle des besoins en services sociaux. Dans les deux cas, ils se sont servis de listes d'annuaire disponibles dans le commerce que l'on a combinées à l'échantillonnage régional des ménages. D'après l'analyse de leurs résultats, ils recommandent l'utilisation de plans à base de sondage double afin de profiter de l'efficacité relative de l'interview téléphonique et d'éliminer les biais inhérents à l'utilisation d'annuaire comme base de sondage. Waksberg, Brick, Shapiro, Flores-Cervantes et Bell (1997) décrivent l'utilisation combinée de l'échantillonnage par CA et de l'échantillonnage aréolaire pour la US National Survey of American families durant laquelle on s'est concentré, en particulier, sur la population à faible revenu. On a remis aux ménages ne possédant pas le téléphone repères durant le dépistage aréolaire des téléphones cellulaires pour leur permettre de répondre aux interviews téléphoniques, donc de ne pas devoir former les intervieweurs chargés des questions de filtrage dans le cas d'un questionnaire non téléphonique (Cunningham, Berlin, Meader, Molloy, Moore et Rajanen 1997).

(1979). Smith et Frazier (1993) comparent les méthodes originales et modifiées, au moyen de données recueillies dans le cadre du California Behavioral Risk Factor Surveillance System. Les résultats indiquent que la méthode modifiée accélère la collecte des données, ce qui produit un échantillon de plus grande taille au même prix. Cet avantage compense les effets plus importants du plan de sondage de cette méthode.

Le recours à la stratification et à la répartition non proportionnelle pour améliorer les « taux d'appels couronnés de succès », proposés par Palit (1983) est une autre variante de la méthode de base de Mitofsky-Waksberg. Palit et Blair (1986) procèdent à une évaluation des autres traitements des numéros de téléphone ne produisant pas de réponse applicables au plan de sondage de Mitofsky-Waksberg. Burke, Morganstein et Schwartz (1981) étudient la détermination optimale des paramètres pour la méthode de Mitofsky-Waksberg et Casady et Lepkowski (1991, 1993), ainsi que Tucker, Casady et Lepkowski (1992) s'intéressent à la répartition optimale pour la version stratifiée. D'autres problèmes ayant trait à la répartition produisant le coût minimal sont examinés par Palit (1983) et par Mason et Immerman (1988).

3.1.4 Méthodes assistées par liste

Bien que les méthodes de CA permettent d'éviter le sous-dénombrement inhérent aux annuaires à cause des numéros non publiés, elle continue de poser le problème fondamental du sous-dénombrement dû aux ménages non abonnés au téléphone (pour plus de détails, consulter la section 3.3). En outre, le manque de données auxiliaires (dont les données géographiques), que fournissent fréquemment les listes, cause des inefficacités, même dans le cas des modifications les plus perfectionnées de la méthode fondamentale exposées plus haut. Par conséquent, on a recherché d'autres méthodes en vue de combiner les échantillons obtenus par CA à des échantillons obtenus d'après des listes et des annuaires. L'un des premiers efforts dans cette direction, qui a été proposé par Stock (1962) et élaboré par Sudman (1973), se fonde sur le remplacement des deux derniers chiffres des numéros de téléphone sélectionnés à partir d'un annuaire par des chiffres sélectionnés au hasard. Hauck et Cox (1974) ont appliqué la méthode pour procéder à une étude méthodologique des effets du mode de sélection lors du filtrage d'une sous-population particulière. Une version simplifiée, appelée communément méthode de l'« unité en plus », consiste à remplacer chaque numéro de téléphone échantillonné d'après un annuaire par le numéro obtenu en ajoutant une unité (ou un autre nombre, ce qui donne alors la méthode du « nombre en plus»). En principe, cette mesure permet d'éviter le biais dû aux numéros non publiés. Étant donné sa simplicité, elle a eu beaucoup de succès auprès des spécialistes des études de marché. Cependant, plusieurs

II présente aussi des comparaisons de coûts et certaines modifications qui permettent de surmonter des problèmes pratiques. Sa méthode réduit l'incertitude quant à la catégorie de numéro composé en cas de non-réponse, ainsi que le problème de l'épuisement des numéros résidentiels dans une UPE.

Un grand nombre de généralisations et de modifications supplémentaires de la méthode fondamentale de Mitofsky-Waksberg ont été proposées. Plusieurs visent à réduire le fardeau du filtrage des interviews et à améliorer le contrôle de la taille de l'échantillon auprès duquel a lieu le premier contact. Ainsi, Hogue et Chapman (1984) proposent de déterminer des nombres limites d'appels en se fondant sur une estimation de la probabilité qu'une UPE soit « peu peuplée », c'est-à-dire qu'elle ne contient qu'une faible proportion de numéros résidentiels, et d'établir une méthode optimale d'établissement du nombre limite en tenant compte du coût et de la variance. Alexander (1988) considère deux types de règles d'exclusion pour limiter le filtrage en vue d'une interview pour les préfixes pour lesquels la densité de numéros résidentiels est faible. Une « règle de croissance » met un terme au processus aussitôt qu'un nombre préétabli d'appels, c_i , ont été faits et que moins de i résidences ont été dépistées, où $\{c_i\}$ est une série croissante dans i . Une « règle de décroissance » met un terme au processus quand i résidences ont été dépistées et sont évaluées dans le cas d'un modèle simple.

Lepkowski et Groves (1986a) proposent un plan d'échantillonnage à deux phases fondé sur l'appariement des préfixes sélectionnés à la première étape de la méthode de Mitofsky-Waksberg aux numéros d'un annuaire commercial pour obtenir le nombre de numéros de téléphones publiés pour chaque préfixe sélectionné. Puis, les préfixes sont répartis en deux strates – une strate à densité faible qui ne contient aucun numéro publié, ou uniquement un petit nombre de ces numéros, et une strate à densité élevée. Puis, on applique le plan d'échantillonnage de Mitofsky-Waksberg à la strate à faible densité et on sélectionne les numéros de téléphone avec probabilité proportionnelle au nombre de numéros de téléphone publiés dans la strate à forte densité.

Brick et Waksberg (1991) proposent d'utiliser un nombre fixe de numéros de téléphone à la deuxième étape, afin d'éviter tout bonnement l'échantillonnage séquentiel, ce qui simplifie l'opération. Le plan de sondage, proposé au départ par Waksberg (1984), n'est toutefois pas autopondéré et produit un léger biais et une augmentation de la variance. Brick et Waksberg (1991) énoncent des points à prendre en considération pour faire un choix entre les plans original et modifié de Mitofsky-Waksberg. Pour l'une des premières applications de la méthode modifiée de Mitofsky-Waksberg à la collecte de données sur les attitudes à l'égard de la santé, qui semble résulter d'un effort erroné d'application de la méthode originale, consulter Cummings

sélectionnée, deux chiffres finals sont sélectionnés au hasard. Le numéro à 10 chiffres résultant est composé et, s'il n'est pas celui d'une résidence (conformément à la définition adoptée pour l'enquête), l'UPB est supprimée de l'échantillon. S'il s'agit d'une résidence, on sélectionne un échantillon aléatoire simple (sans remise) de k numéros résidentiels supplémentaires. La méthode de sélection des UPB se poursuit jusqu'à ce que l'on ait sélectionné un nombre établi, m , d'UPB. Il est facile de voir, si l'on suppose que le nombre de numéros résidentiels dans chaque UPB sélectionnée, P_i , est au moins égal à k , que l'effectif total de l'échantillon de ménages ayant un numéro de téléphone résidentiel est $m(k+1)$ et que l'échantillon final est un échantillon de la population complète de ménages ayant un numéro de téléphone résidentiel sélectionné avec probabi-

lité égale. Wakseberg (1978) montre que, si nous représentons par $\pi = (\sum_{i=1}^M P_i^2) / (NM)$ la proportion de numéros résidentiels dans la population et par t la proportion d'UPB sans numéro résidentiel (c'est-à-dire pour lesquelles $P_i = 0$), alors le nombre prévu total d'appels est donné par $m[1 + (1-t)k]/\pi$, en supposant que $P_i' \geq k+1$ pour toutes les UPB qui comptent au moins un numéro résidentiel. On peut abandonner la dernière hypothèse si l'on regroupe les UPB de sorte que la contrainte soit vérifiée pour chaque groupe ou si l'on applique des coefficients de pondération inégaux. On obtient les valeurs optimales des paramètres du plan de sondage dans le cas d'une fonction simple de coût et on étend la méthode au traitement d'enquêtes répétées. L'avantage principal de la méthode tient à la réduction du nombre prévu d'appels qu'il faut faire pour obtenir une taille donnée, efficace, d'échantillon, particulièrement si t , c'est-à-dire la proportion d'UPB sans numéro de téléphone résidentiel, est supérieur à 0,5. Groves (1977) publie, pour une étude nationale, des données indiquant que la valeur de t est d'environ 0,65. Il faut mettre en balance cet avantage et l'augmentation de la variance due à la mise en grappe. Néanmoins, si l'on tient compte des coûts, des calculs illustratifs pour des valeurs types des paramètres montrent que la réduction des coûts est de l'ordre de 20 % à 40 %.

Le principal inconvénient opérationnel de la méthode est dû à sa nature séquentielle qui rend difficile son application manuelle. Par contre, l'opération séquentielle ne pose aucun problème si le procédé de sélection est entièrement automatisé. Telle qu'elle est décrite plus haut, la méthode pose d'autres problèmes, que, dans la plupart des cas, de simples modifications permettent de surmonter. Si l'on suppose que l'on ne possède aucun renseignement a priori sur le nombre de ménages possédant un numéro de téléphone, on ne connaît pas les probabilités de sélection, mais on peut estimer la valeur de p à partir de l'échantillon. La nécessité pratique d'introduire une règle d'arrêt pour limiter le nombre d'appels sans répondant ou pour lesquels le

répondant refuse de répondre, même s'il s'agit d'un numéro résidentiel, sous-entend que l'on ne peut appliquer la méthode strictement telle qu'elle est conçue, ce qui pourrait biaiser les résultats. Il est possible de contourner le problème des ménages qui possèdent plusieurs numéros de téléphone si l'on obtient des renseignements corrects sur le nombre de lignes distinctes, mais la pondération corrective qui est nécessaire réduit la simplicité de la pondération uniforme. Dans certains cas, il est possible d'obtenir les noms et les adresses par appartement des numéros obtenus par CA à des listes d'adresses, si bien que l'on peut envoyer un avis préalable à une partie des répondants éventuels. Cependant, la procédure est complexe et les difficultés que pose l'envoi d'un avis aux répondants (communes à toutes les méthodes de CA) rendent sa prise en considération difficile pour certains bureaux officiels de la statistique.

3.1.3 Modifications de la méthode de Mitofsky-Wakseberg et d'autres méthodes de compositions aléatoires

Certains inconvénients de la méthode de base peuvent être surmontés grâce à la généralisation proposée par Pothoff (1987a, 1987b). La méthode se fonde sur la définition d'un ensemble de numéros de téléphone favorable d'un ensemble de numéros résidentiels. On définit, comme dans le cas de la méthode de Mitofsky-Wakseberg, ou d'un ensemble plus général qui inclut tous les numéros résidentiels – par exemple, l'ensemble de numéros pour lesquels on obtient une tonalité (y compris le signal occupé, les messages enregistrés et les téléphonistes). La première étape correspond à un échantillonnage aléatoire simple d'un nombre déterminé, m , d'UPB. À partir de chaque UPB sélectionnée, on effectue un nombre déterminé d'appels, c , et, pour chaque appel, on détermine si le numéro est favorable ou non. Toute UPB pour laquelle les c numéros sélectionnés sont défavorables est rejetée. Les UPB retenues sont considérées comme étant de type I si elles ne comportent qu'un seul numéro favorable et de type II si elles comptent au moins deux numéros favorables. La deuxième étape consiste à sélectionner et à composer kc numéros provenant des UPB de type I et $k(c-1)$ numéros provenant des UPB de type II, où k est un nombre entier. Pour chaque numéro composé, on détermine si l'unité est résidentielle ou hors du champ d'observation, et l'on s'efforce de réaliser une interview dans le cas de toutes les unités résidentes. Pour les UPB de type I, un segment séquentiel supplémentaire sélectionne des numéros de téléphone supplémentaires qui sont composés jusqu'à ce que l'on obtienne un total de k numéros favorables. Une tentative d'interview est faite pour chaque numéro favorable composé dans le cadre du segment séquentiel. Pothoff (1987a) montre que, dans certaines conditions, tous les numéros de téléphone résidentiels ont la même probabilité de sélection et développe des estimations non biaisées et des rapports estimatifs, ainsi que leurs variances.

suffixes aléatoires à quatre chiffres à des préfixes connus pour réaliser une enquête locale. Par la suite, Eastlack et Assael (1966) et Glasser et Metzger (1972) ont amélioré cette méthode d'échantillonnage de base et lui ont donné une portée nationale grâce à la création de « banques valides » de numéros d'après les renseignements fournis par les compagnies de téléphone.

Jusqu'à récemment, la CA n'était utilisée, en grande partie, qu'aux États-Unis et au Canada. En effet, Sykes et Collins (1987) indiquent que les enquêtes téléphoniques étaient encore rares au Royaume-Uni à la fin des années 1980, principalement à cause de la faible pénétration du téléphone. Plus précisément, on recourait rarement à la CA, notamment à cause du manque d'uniformité de la longueur des numéros de téléphone à l'époque. Toutefois, récemment, étant donné la plus grande pénétration du téléphone au Royaume-Uni, de l'ordre de 96 % à la fin des années 1990, et la normalisation des numéros de téléphone qui comptent maintenant dix chiffres, les enquêtes par CA sont plus fréquentes – à cet égard, consulter, par exemple, Collins (1999) et Nicolaas, Lynn et Lound (2000). Pareillement, Gabler et Haeder (2000) indiquent qu'une méthode par CA, modifiée pour tenir compte de la longueur variable des numéros de téléphone (de 6 à 11 chiffres) est désormais la méthode normalisée d'enquête téléphonique en Allemagne.

Mitofsky (1970) a proposé le premier une méthode d'échantillonnage à deux degrés par CA pour parer à l'inefficacité des méthodes de CA élémentaires due au fait qu'il faut composer un grand nombre de numéros qui ne produisent pas d'interview (numéros non en service et numéros non résidentiels). Cette méthode, que Wakseberg (1978) a élaborée par la suite et à laquelle il a donné un fondement théorique solide, a été appelée méthode de Mitofsky-Wakseberg. Cette dernière ou ses variantes sont devenues les principales méthodes d'échantillonnage appliquées aux enquêtes téléphoniques, du moins aux États-Unis.

Elle se fonde sur le fait que les numéros de téléphone résidentiels sont, en général, regroupés en séries de numéros consécutifs ou dans des banques de numéros ayant les mêmes « premiers chiffres ». Pour les États-Unis, la valeur de r est habituellement fixée à huit (pour les numéros de téléphone à 10 chiffres, y compris l'indicateur régional), de sorte que la taille de chaque banque ou grappe (UPF-unité primaire d'échantillonnage) soit $N = 100$. On suppose que la compagnie de téléphone peut fournir la liste de tous les préfixes opérationnels (indicateur régional plus les trois premiers chiffres du numéro), c'est-à-dire ceux auxquels a été attribué un numéro résidentiel. On ajoute à chaque numéro à six chiffres de cette liste toutes les combinaisons possibles de deux chiffres, afin d'obtenir une base de sondage de numéros à huit chiffres qui représente les M UPF qui forment la population. Les UPF de l'échantillon sont sélectionnées au hasard, de façon consécutive, à partir de cette base de sondage (avec remise) et, pour chaque UPF

pas à la population observée est incluse dans la base de sondage. Cette situation pourrait tenir au fait que, souvent, les numéros qui ne sont plus en service continuent de figurer dans l'annuaire, au fait que les numéros commerciaux ne sont pas toujours clairement désignés en tant que tels ou à d'autres cas non reconnus d'indéterminabilité. Le compte multiple a lieu lorsque la même unité est représentée plus d'une fois dans la base de sondage et que ce fait n'est pas reconnu. On peut habituellement repérer les comptes multiples durant l'échantillonnage si les entrées pour un même ménage sont énumérées consécutivement, mais on ne le peut pas si elles apparaissent séparément (par exemple, sous divers noms de famille). Si le compte multiple est confirmé durant l'interview (par exemple, en obtenant des renseignements sur le nombre de lignes raccordées dont dispose le ménage ou le nombre de numéros publiés dans l'annuaire) l'application de coefficients de pondération appropriés permet de résoudre le problème. Alors que le suréchantillonnage est surmontable, à un certain prix, le sous-échantillonnage ne l'est pas, ce qui indique qu'il faut recourir à des bases de sondage plus représentatives que ne le sont les annuaires. Les listes produites par les entreprises commerciales, ordinairement pour des raisons de marketing, sont devenues une solution fréquente pour remplacer le traditionnel annuaire téléphonique (produit en général par la compagnie qui fournit le service téléphonique dans la région). Ces listes peuvent correspondre aux annuaires de la ville, être produites d'après les listes municipales d'adresses et complétées par les numéros provenant d'annuaires ou d'autres sources, les listes d'adresses des compagnies de téléphone ou les listes nationales principales d'adresses, telles que celles produites par Donnelly Marketing, Inc. aux États-Unis (Lepkowski 1988). Ces listes fournissent des données auxiliaires importantes, comme les données géographiques provenant du Recensement de la population et du logement et d'autres sources. En général, elles ne permettent pas d'éliminer le biais dû aux numéros non publiés et leur coût peut être élevé. Leur utilisation améliore parfois la variance d'échantillonnage, grâce aux données auxiliaires qui permettent d'élaborer un plan de sondage plus efficace. Eventuellement, les listes utilisées par les services d'urgence pour déterminer l'emplacement physique des personnes qui appellent pourraient servir de bases de sondage, mais les organismes d'enquêtes non gouvernementaux auraient de la difficulté à les obtenir.

3.1.2 Composition aléatoire – La méthode de Mitofsky-Wakseberg

L'application de la composition aléatoire (CA) pour réaliser les enquêtes téléphoniques est devenue un moyen très populaire, particulièrement aux États-Unis. Ces méthodes de CA se fondent sur une base de sondage qui englobe tous les numéros de téléphone possibles. L'idée a été proposée au départ par Cooper (1964), qui a ajouté des

L'Enquête sur la population active au Canada (Drew, Choudhry et Hunter 1988). Dans ces cas, l'échantillonnage a pour base de sondage une liste générale à laquelle sont ajoutées des renseignements sur les numéros téléphoniques, et l'élaboration du plan de sondage ne tient compte d'aucune caractéristique particulière de l'usage du téléphone. Il en est de même pour les enquêtes téléphoniques « pures » réalisées auprès de populations particulières, comme les médecins, pour lesquelles on possède la liste complète des membres de la population avec les numéros de téléphone qui peut alors servir de base de sondage (voir, par exemple, Gunn et Rhodes 1981). Un autre exemple est celui où l'on recourt à l'interview téléphonique lors des cycles de suivi d'une enquête par panel pour laquelle le premier contact prend la forme d'un interview sur place. Par exemple, dans le cas de l'Enquête sur la population active d'Israël, la première prise de contact se fait lors d'une visite sur place et les deuxième et troisième cycles sont réalisés par téléphone auprès des ménages disposés à participer à l'enquête de cette façon (Nathan et Elivav 1988). Une méthode apparemment utilisée récemment lors d'une étude pilote réalisée pour la US National Study of Health and Activity (Maffeo, Frey et Kalton 2000), consiste à tirer un échantillon régional, à obtenir dans la mesure du possible les numéros de téléphone en vue d'une interview téléphonique et de procéder à l'interview sur place pour ceux dont on n'a pas obtenu le numéro et pour ceux qui n'ont pas répondu à l'enquête téléphonique.

L'annuaire le plus facile à obtenir et le moins coûteux à utiliser comme base de sondage pour les enquêtes téléphoniques est, naturellement, l'annuaire téléphonique proprement dit, ou une version modifiée de cet annuaire. Au départ, on se servait de la version imprimée de l'annuaire, tandis qu'aujourd'hui, on se sert de la version électronique. Les principaux inconvénients de l'utilisation de l'annuaire téléphonique comme base de sondage, c'est-à-dire le sous-dénombrement, le surdénombrement, ont été bien décrits. Le sous-dénombrement, qui est de loin le défaut le plus grave, inclut les ménages qui ne possèdent pas le téléphone, ainsi que ceux qui choisissent d'avoir un numéro de téléphone non publié ou ceux dont le numéro n'a pas encore été inséré dans l'annuaire. Nous traiterons à la section 3.3 du biais dû aux ménages qui ne possèdent pas le téléphone, lequel ne dépend naturellement pas de la base de sondage utilisée.

La proportion de numéros de téléphone non publiés varie considérablement selon le pays et le genre d'emplacements, ainsi que d'autres variables du ménage. Sykes et Collins (1987) publient un taux de numéros non publiés de 4 % pour les Pays-Bas et de 12 % pour le Royaume-Uni. Fréjean, Panzani et Tassi (1990) estiment à 14 % le taux de numéros non publiés en France et, aux États-Unis, durant les années 1970, on estimait que ce taux excédait 17 % à 19 % (Blankenship 1977b; Glasser et Metzger 1975). Selon Rich (1977), dans la région de la Californie desservie par Pacific Telephone, le taux de numéros de téléphone non publiés

(excluant les numéros involontairement non publiés) est passé de 9 % en 1964 à 28 % en 1977. En outre, en Californie, environ 5 % de numéros de téléphone résidents seraient involontairement non publiés (attribués après la publication de l'annuaire). Selon des études plus récentes, le taux de numéros non publiés est considérablement plus élevé. Ainsi, d'après Genesys (1996), le taux national était de 40 % en 1993 et de 37 % en 1995, calculé d'après des échantillons nationaux de près de 100 000 interviews téléphoniques par CA, et selon Survey Sampling Inc. (1998), en 1997, le taux estimatif national de numéros non publiés aux États-Unis était de 30 %. Une étude à petite échelle réalisée dans la région de Jérusalem (Nathan et Aframian 1996) indique un taux de numéros non publiés de 27 %.

De nombreuses études révèlent des écarts importants entre les caractéristiques des ménages dont le numéro de téléphone est publié et de ceux dont le numéro ne l'est pas, différences qui pourraient être la source d'un biais de couverture en cas d'échantillonnage fondé sur l'annuaire téléphonique. Aux États-Unis, ces différences ont été mises en évidence, par exemple, par une étude de Brunner et Brunner (1971), qui ont observé des écarts hautement significatifs entre les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas, pour toute une série de variables démographiques et socioéconomiques. Leuthold et Schaele (1971) ont observé un taux plus élevé de non publication chez les noirs, les citadins, les jeunes, les personnes qui vivent en appartement, les personnes divorcées ou séparées et les travailleurs du secteur des services. Partiellement, Roslow et Roslow (1972) ont constaté des écarts significatifs entre les parts d'auditoire pour les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas. Glasser et Metzger (1975) ont montré que le taux de numéros non publiés était plus élevé dans l'Ouest, dans les grandes régions métropolitaines, chez les populations non blanches et chez les jeunes. Blankenship (1977b) et Rich (1977) ont noté des écarts hautement significatifs entre les ménages dont le numéro de téléphone est publié et ceux dont le numéro ne l'est pas en ce qui concerne le sexe et l'âge du chef de ménage, la profession, la taille du ménage et le revenu. Au Royaume-Uni, Sykes et Collins (1987) ont observé une plus forte proportion de numéros non publiés chez les jeunes, les personnes les plus pauvres et celles vivant à Londres. Pour la région de Jérusalem, Nathan et Aframian (1996) ont montré que les taux de ménages possédant un téléviseur et de ménages regardant la télévision (pour ceux qui possèdent un téléviseur) étaient plus faibles pour un échantillon sélectionné par CA que pour un échantillon sélectionné d'après un annuaire.

Outre le sous-dénombrement dû aux numéros non publiés, tels que décrits plus haut, les listes d'annuaire présentent les problèmes du surdénombrement, du compte multiple et du manque de renseignements auxiliaires à jour. Le surdénombrement a lieu lorsqu'une unité n'appartenant

de ces enquêtes. Les résultats ont été publiés dans des monographies ou dans des numéros spéciaux de revues scientifiques. En novembre 1987 s'est tenue, à Charlotte, en Caroline du Nord, une conférence importante sur la méthodologie des enquêtes téléphoniques qui a été suivie par la publication du compte rendu sous la direction de Groves, Biemer, Lyberg, Masssey, Nicholls et Wakseberg (1988) et d'un numéro spécial de la revue *Journal of Official Statistics*, sous la direction de Groves et Lyberg (1988b). La conférence sur la technologie des enquêtes assistées par ordinateur tenue à Berkeley au printemps 1981 (Freeman et Shanks 1983) portait principalement sur les enquêtes téléphoniques. L'ITAO a été le sujet principal de la conférence internationale sur la collecte de données d'enquête assistée par ordinateur InterCASIS 96 tenue à San Antonio, Texas, en décembre 1996 (Couper, Bethlehem, Baker, Clark, Martin, Nicholls et O'Reilly 1998) et de la troisième conférence internationale de l'ASC qui s'est tenue à Edinbourg en septembre 1999 (Banks, Westlake 1999).

Les auteurs susmentionnés fournissent des bibliographies détaillées comportant plusieurs centaines d'articles, de même que Khurshid et Sahai (1995), qui couvrent la période jusqu'à 1991, et *Survey Research Center* (2000), qui donne une mise à jour des bibliographies antérieures en ce qui concerne les plans d'échantillonnage pour les enquêtes téléphoniques auprès des ménages jusqu'à 2000. La partie qui suit est consacrée à l'examen du développement des méthodes d'enquêtes téléphoniques après des menages durant les 25 dernières années en se concentrant sur le plan d'échantillonnage, l'estimation, la couverture, la non-réponse et l'évaluation de la qualité des données.

3.1 Plan d'échantillonnage et estimation

La méthode d'échantillonnage appliquée aux enquêtes téléphoniques se fonde sur les principes généraux de l'échantillonnage. Son adaptation au contexte particulier des enquêtes téléphoniques a trait principalement à la base de sondage utilisée. Donc, nous adoptons la classification proposée par Lepkowski (1988) pour les méthodes d'échantillonnage téléphonique, selon la base de sondage utilisée – aléatoire de numéros de téléphone (CA) et méthodes combinées (assistées par liste et base de sondage double).

3.1.1 Méthode d'échantillonnage axée sur une liste

Comme nous l'avons mentionné plus haut, les premières enquêtes téléphoniques étaient réalisées auprès d'échantillons sélectionnés d'après des listes. Il s'agissait souvent d'enquêtes à mode mixte où l'on recourait à l'interview téléphonique pour compenser la non-réponse lors des interviews sur place ou pour réaliser un suivi. Ces scénarios dits de « interview téléphonique à chaud » ont été appliqués à la Current Population Survey aux États-Unis et à

Des centaines d'articles scientifiques ayant trait à un large éventail d'aspects des enquêtes téléphoniques ont été publiés durant cette période. Plusieurs ouvrages généraux sur le sujet ont paru – Blankenship (1977), Groves et Kahn (1979), Frey (1989) et Lavrakas (1993). Un certain nombre de conférences ont été consacrées à la méthodologie des enquêtes téléphoniques ou ont traité d'aspects particuliers

Durant le dernier quart de siècle, l'enquête téléphonique a définitivement fait ses preuves. Lyberg et Kasprzyk (1991) soutiennent qu'elle est devenue le mode principal de collecte de données dans les pays où la pénétration du téléphone est grande.

3. PROGRÈS RÉCENTS DANS LE DOMAINE DES ENQUÊTES TÉLÉPHONIQUES

l'ITAO – par exemple, le système A&S/CAT™ (Dutka et Frankel 1980). Les organismes universitaires d'enquête par sondage leur ont emboîté le pas rapidement grâce au premier système mis au point à l'UCLA et à Berkeley pour la grande enquête sur l'incapacité en Californie basée sur l'ITAO (Shanks, Nicholls et Freeman 1981; et Shanks 1983). Un autre projet précoce de mise au point d'un système d'ITAO par un organisme universitaire d'enquête, selon une approche différente fondée sur des micro-ordinateurs, a été celui de l'Université du Wisconsin (Palit 1980; Palit et Sharp 1983). En Europe, les premiers organismes de recherche par sondage qui ont utilisé l'ITAO ont été le Social and Community Planning Research (SCPR, appelé aujourd'hui National Centre for Social Research) au Royaume-Uni (Sykes et Collins 1987) et l'Université d'état d'Utrecht, aux Pays Bas (Dekker et Dom 1984). L'adoption des systèmes d'ITAO par les bureaux officiels de la statistique a été plus lente. Aux États-Unis, elle a débuté en 1982 au Censur Bureau (Nicholls 1983) et au National Agricultural Statistics Service (Tortora 1985); parallèlement, elle a débuté la même année au bureau de la statistique des Pays Bas (1987). En 1987, selon un sondage mené auprès d'un échantillon (non probabiliste) de 27 organismes d'enquête (18 aux États-Unis et 9 ailleurs), presque tous utilisaient l'ITAO pour certaines ou toutes leurs enquêtes téléphoniques (Berry et O'Rourke 1988). Le rapport du Federal Committee on Statistical Methodology (1990) indique que le nombre estimatif d'installations d'ITAO dans le monde à la fin des années 1980 était supérieur à 1 000, et en 1988, le gouvernement américain collaborait avec 51 centres d'ITAO. Il convient de souligner que le développement de l'ITAO a fait rapidement part d'un mouvement plus général vers la collecte de données d'enquête assistée par ordinateur (CDEAO), qui englobe aussi l'interview sur place assistée par ordinateur (IPAO) et l'auto-interview assistée par ordinateur (AIAO) (Nicholls 1988). Pour un historique plus complet de la mise au point de l'ITAO et de la CDEAO en général, consulter Couper et Nicholls (1998).

prendre catégoriquement position. Lors d'une comparaison de l'envoi d'un questionnaire par la poste, de l'interview téléphonique et de l'interview sur place, Wiseman (1972) observe un effet de mode d'interview dans le cas de questions délicates (sur l'avortement et la contraception). Toutefois, c'est pour l'envoi d'un questionnaire par la poste et l'interview personnelle (par téléphone ou sur place) que la différence est la plus importante.

Les résultats de plusieurs études empiriques plus rigoureuses ont apaisé rapidement nombre des craintes susmentionnées. Ainsi, lors d'une expérience contrôlée bien conçue, Hochstim (1967) compare la collecte des données quand le mode principal est l'envoi d'un questionnaire par la poste, l'interview téléphonique ou l'interview sur place. L'étude donne des preuves convaincantes que les trois méthodes de collecte des données sont presque interchangeables si l'on s'en tient au taux de réponse, à la complétude de la réponse, à la comparabilité des résultats et à la validité des réponses. La différence importante entre les divers modes de collecte a trait au coût et joue manifestement en faveur de l'enquête par la poste ou par téléphone. Par ailleurs, un petit essai réalisé par Colomoto (1965) sur des échantillons d'une population de médecins ne révèle aucun écart significatif entre les réponses obtenues par téléphone et en personne. Janofsky (1971) constate que les répondants sont aussi disposés à exprimer leurs sentiments concernant des questions de santé par téléphone qu'en personne. L'étude de validation bien conçue menée par Locander, Sudman et Bradburn (1976) sur les effets du mode de collecte en cas de questions délicates ne révèle aucune différence significative entre les biais de réponse associés à l'interview téléphonique et à l'interview sur place. Enfin, au moyen d'une petite expérience sur le terrain minutieusement contrôlée, Rogers (1976) a testé les effets de diverses stratégies d'interview sur la qualité des réponses et sur les résultats sur le terrain, lors d'une enquête comportant une gamme de questions complexes sur les attitudes, les connaissances et les caractéristiques personnelles. De nouveau, l'étude montre que les données obtenues par téléphone sont de même qualité que celles obtenues lors d'interviews sur place. Groves et Kahn (1979) ont réalisé une étude nationale importante en vue de comparer l'interview téléphonique et l'interview sur place. Leur analyse approfondie des résultats de grandes enquêtes omnibus réalisées en appliquant les deux méthodes de collecte par University of Michigan Survey Research Center a fourni des données importantes sur la qualité des données qui ne témoignent d'aucun effet significatif du mode de collecte. Cette étude et d'autres qui ont été les précurseurs des études systématiques des effets du mode de collecte réalisées durant les années 1990 (décrites plus loin) ont contribué à la légitimation des enquêtes téléphoniques en tant que méthode type de collecte des données.

Au départ, l'utilisation du téléphone pour réaliser les enquêtes par sondage se fondait surtout sur des échantillons sélectionnés à partir de bases de sondages générales, comme

les annuaires téléphoniques, ou à partir de bases de sondages particulières pour les petites sous-populations. Vers la fin des années 1960, on avait pris conscience du taux croissant de numéros de téléphone non publiés et des différences importantes entre les numéros de téléphone publiés et ceux dont le numéro ne l'était pas (pour plus de précisions, consulter la section 3.1.1). La méthode d'échantillonnage par composition aléatoire (CA), introduite pour la première fois par Cooper (1964) et améliorée par la suite par Eastlack et Assael (1966) et par Glasser et Metzger (1972), représente un progrès important qui a permis de surmonter le problème. Une des inefficacités inhérentes aux trois méthodes fondamentales de CA tenait au grand nombre de numéros composés qui ne produisaient pas d'interview (numéros non en service ou non résidentiels). Mitofsky (1970) a été le premier à proposer une méthode d'échantillonnage par CA à deux degrés pour contourner ce problème; par la suite, Waksberg (1978) l'a élaborée et lui a donné un support théorique solide. L'introduction de cette méthode, qui allait devenir la méthode de Mitofsky-Waksberg, a beaucoup contribué à la généralisation des enquêtes téléphoniques durant les années 1960 et 1970.

Enfin, les progrès réalisés durant les années 1960 et 1970 dans le domaine des télécommunications et de l'automatisation ont aussi donné un avantage à l'enquête téléphonique. L'accès universel à la composition automatique des numéros interurbains a permis de réaliser plus facilement des enquêtes nationales à partir d'un seul centre téléphonique ou à partir d'un petit nombre de centres d'interview avec tous les avantages d'un contrôle et d'une administration centralisés. Cependant, l'élément qui a stimulé le plus l'expansion des enquêtes téléphoniques est, sans aucun doute, l'introduction de l'interview téléphonique assistée par ordinateur (ITAO) durant les années 1970. Cet effet important tient à la simplicité de l'ITAO pour réaliser des enquêtes téléphoniques et aux possibilités qu'elle offre en ce qui concerne l'automatisation de nombreuses tâches importantes connexes à l'interview (comme la composition, l'établissement des calendriers de

L'une des premières applications de l'ordinateur à l'interview téléphonique a pris la forme d'une expérience de laboratoire axée sur l'utilisation d'un ordinateur muni de plusieurs postes de travail conçus pour recueillir des renseignements subjectifs (Shure et Meeker 1970). Le numéro spécial de *Sociological Methods and Research* (Freeman et Shanks 1983), publié après la conférence de Berkeley sur les techniques d'enquête assistée par ordinateur qui s'est tenue au printemps 1981 décrit bien les débuts de l'ITAO. Les organismes spécialisés dans les études de marché ont été les premiers à adopter les systèmes d'ITAO pour leurs opérations courantes. En 1972, Chilton Research Services avait déjà mis au point et utilisait régulièrement le Survey Response Processor (Fink 1983). D'autres organismes commerciaux d'enquête qui utilisaient des systèmes différents se sont vite rendus compte des avantages de

progrès technologiques courants et futurs du secteur des communications sur les pratiques d'enquête et leurs conséquences méthodologiques.

2. LES DÉBUTS DES ENQUÊTES TÉLÉPHONIQUES

Dans la section qui suit, nous examinons, brièvement et globalement, les débuts de l'utilisation du téléphone pour réaliser les enquêtes, afin de placer dans son contexte le développement des méthodes de téléenquête que nous décrivons plus tard. Le sujet est abordé de façon plus détaillée et plus complète dans plusieurs traités et articles, dont Biankeship (1977a), Groves, Biemer, Lyberg, Massey, Nicholls et Waksberg (1988), Frey (1989), Lavrakas (1993), Casady et Lepkowski (1998, 1999) et Dillman (1978, 2000).

Le téléphone a commencé à être utilisé pour réaliser des sondages dans les années 1930, en général comme moyen supplémentaire de collecte. Certains attribuent au moins partiellement, mais à tort, au sous-développement téléphonique l'échec désastreux du sondage du *Literary Digest* qui avait prédit la victoire écrasante de Landon sur Roosevelt en 1936 (Katz et Cantil 1937; Payne 1956; Perry 1968). En fait, le sondage se fondait sur un questionnaire envoyé par la poste et, bien que l'on ait utilisé comme base de sondage des listes téléphoniques (conjugées à des listes d'entretiens de véhicules à moteur), il semble que l'échec soit attribuable davantage à la non-réponse qu'à la sous-représentation de la base de sondage (Bryson 1976; Squire 1988; Cahalan 1989).

Les premiers rapports sur la réalisation d'enquêtes par téléphone concernaient, pour la plupart, le domaine de la santé publique ou les applications d'étude de marché. Nombre de ces enquêtes s'appuyaient sur la combinaison d'interviews téléphoniques à d'autres modes de collecte et, dans certains cas, comportaient des comparaisons empiriques de taux de réponse ou de résultats, en vue d'évaluer les effets du mode de collecte des données. Par exemple, Cunningham, Westerman et Fischhoff (1956) et Bonnet (1961) décrivent l'utilisation d'enquêtes téléphoniques pour des études de suivi du traitement des patients, tandis que Fry et McNaire (1958) la décrivent dans le cas du suivi national d'une enquête par la poste en vue de recueillir les opinions du personnel hospitalier – utilisation qui a produit dans l'un et l'autre cas un taux de réponse élevé. Mitchell et Rogers (1958) ont recouru à l'interview téléphonique pour réaliser auprès des ménages ayant le téléphone une enquête sur la consommation de produits laitiers et ont comparé les résultats à ceux obtenus auprès d'un échantillon témoin de ménages ne possédant pas le téléphone. Cahalan (1960), quant à lui, compare les résultats d'interviews téléphoniques à ceux d'interviews sur place réalisées pour évaluer le lectorat des journaux et obtient des résultats favorables. Grâce à une étude téléphonique comparative de

la remémoration de la publicité et de l'utilisation des produits, Eastlack (1964) montre que le protocole rigoureux de rappels produit des résultats plus exacts qu'une méthode sans rappel. Coombs et Freedman (1964) mentionnent un taux de réponse élevé (92 %) dans le cas d'une enquête longitudinale sur la fécondité complétée par des interviews sur place. Sudman (1966) décrit plusieurs autres applications du téléphone pour réaliser les enquêtes, y compris la prise préalable d'un rendez-vous et le dépistage des populations rares, qui augmentent les taux de coopération et permettent de réduire les coûts.

C'est à la fin des années 1960 que les enquêtes téléphoniques ont vraiment pris leur essor, grâce à plusieurs faits nouveaux. En premier lieu, la croissance rapide de la pénétration du téléphone en Europe de l'Ouest et en Amérique du Nord a permis d'adopter l'interview téléphonique comme mode principal de collecte des données. Aux États-Unis, la proportion de ménages équipés du téléphone atteignait 88 % en 1970 (Massey et Bowman 1988) et ce niveau de pénétration a été atteint un peu plus tard dans les pays d'Europe de l'Ouest, en Australie et en Nouvelle-Zélande (Trewin et Lee 1988). Parallèlement à l'augmentation rapide de la pénétration du téléphone dans de nombreux pays, à la fin des années 1960, les enquêteurs ont constaté une baisse importante des taux de réponse et éprouvé des difficultés à communiquer sur place avec les répondants pour recueillir les données. Cette situation les a poussés à envisager sérieusement le passage aux enquêtes téléphoniques en vue de réduire les coûts et d'obtenir un taux plus élevé de coopération. Ce sont les organismes commerciaux et universitaires de sondage qui ont adopté le plus rapidement l'interview téléphonique; les bureaux gouvernementaux de la statistique ont mis plus de temps à le faire. Par exemple, selon Federal Committee on Statistical Methodology (1984) en 1981, 11 % seulement des enquêteurs de l'administration fédérale américaine compaient une composante d'interview téléphonique, dans la plupart des cas combinée à d'autres modes de collecte.

Au début, l'interview téléphonique suscitait des appréhensions, même quand elle servait uniquement de mode complémentaire de collecte, car on craignait que le taux de non-réponse soit élevé et qu'un biais de réponse, considéré comme inhérent aux interviews non réalisées sur place, entâche les données. Les résultats des premières enquêtes téléphoniques semblaient, dans certains cas, confirmer ces craintes. Par exemple, une étude de la réception d'un dépliant réalisée par Larson (1952) suscite de sérieux doutes, fondés sur des interviews sur place subséquentes, quant à la validité des réponses téléphoniques. Parallèlement, lors d'une enquête sur l'amélioration d'un service à la clientèle, Oakes (1954) obtient un taux de réponse beaucoup plus faible dans le cas de l'interview téléphonique que dans celui de l'interview sur place. Dans le cadre d'une enquête attitudinale sur les finances des consommateurs, Schmiedeskamp (1962) constate que les personnes interviewées par téléphone sont davantage portées à ne pas

Méthodes de téléenquêtes applicables aux enquêtes-ménages — Revue et réflexions sur l'avenir

GAD NATHAN¹

RÉSUMÉ

Nous appelons « téléenquête » les enquêtes pour lesquelles le mode principal ou unique de collecte des données repose sur un moyen électronique de télécommunications y compris le téléphone et d'autres dispositifs technologiques plus avancés, enquêtes par téléphone et, plus en détail, les progrès récents dans les domaines du plan de sondage et de l'estimation, de la couverture et de l'évaluation de la qualité des données. Ces progrès méthodologiques ont fait de l'enquête téléphonique le mode principal de collecte des données dans le domaine des enquêtes par sondage au cours du dernier quart de siècle. D'autres moyens de télécommunication de pointe deviennent rapidement des compléments importants, voire même des concurrents, du service téléphonique fixe et sont déjà utilisés de diverses façons pour réaliser les enquêtes par sondage. Nous examinons leur potentiel pour les opérations d'enquête et l'effet que pourraient avoir les progrès technologiques actuels et futurs dans le secteur des télécommunications sur les pratiques d'enquête et leurs conséquences méthodologiques.

MOTS CLÉS : Enquêtes téléphoniques; enquêtes par Internet; plan d'échantillonnage; non-réponse; couverture.

1. INTRODUCTION

À l'aube du nouveau millénaire, les télécommunications électroniques sont devenues un élément prédominant de presque tous les aspects de la vie contemporaine. Nos enquêtes par sondage ne font pas exception et le recours généralisé au téléphone comme mode principal de communication durant au moins le dernier quart de siècle, a fortement influencé les pratiques d'enquête. En fait, dans le domaine des enquêtes par sondage, l'enquête téléphonique est devenue l'instrument dominant de collecte des données, surtout en Amérique du Nord et en Europe de l'Ouest, qu'il s'agisse d'enquêtes menées auprès de ménages, de particuliers ou d'établissements. D'autres modes de télécommunication de pointe, comme le courrier électronique, Internet, la vidéophonie, la télécopie et la téléphonie mobile, deviennent rapidement des compléments importants, voire même des concurrents, du service téléphonique fixe. Ils sont déjà appliqués de diverses façons aux enquêtes par sondage et nous avons l'intention d'examiner ici les possibilités qu'ils offrent pour les opérations d'enquête et les conséquences méthodologiques de leur utilisation. Par conséquent, nous qualifions de « téléenquête » toute enquête pour laquelle le mode prédominant ou unique de collecte des données repose sur un moyen électronique de télécommunications, y compris le téléphone et d'autres dispositifs technologiques plus avancés. Les enquêtes classiques, fondées sur l'interview sur place, ou les enquêtes par la poste ne sont pas incluses, à moins qu'une de leurs composantes importantes ne s'appuie sur un instrument de télécommunication. Bien que le présent article se concentre sur les enquêtes auprès des particuliers et des ménages, une

Nous nous attachons principalement aux aspects statistiques de la méthodologie des téléenquêtes, tout en reconnaissant qu'elles ne sont pas indépendantes des aspects non statistiques, comme les caractéristiques cognitives de la téléinterview, l'administration des enquêtes et les considérations d'ordre éthique. Dans la section qui suit, nous passons brièvement en revue les débuts des enquêtes téléphoniques, jusqu'à 1978. À la troisième section, nous examinons de façon assez approfondie les faits plus récents dans le domaine du plan d'échantillonnage et de l'estimation, de la couverture et de la non-réponse, ainsi que de l'évaluation de la qualité des données. Enfin, à la quatrième section, nous étudions les effets que pourraient avoir les

Le présent article est rédigé en reconnaissance de la contribution unique de Joe Waksberg à l'élaboration de la méthodologie d'enquête en général et de celle des enquêtes téléphoniques en particulier. Il est généralement admis aujourd'hui que son article innovateur (Waksberg 1978) a frayé la voie à l'application généralisée et efficace de la composition aléatoire aux enquêtes téléphoniques et représente un jalon dans le développement de la méthodologie des téléenquêtes. Conjointement aux nombreux articles qu'il a publiés par la suite, ce document a eu une profonde influence théorique et pratique sur la méthodologie des enquêtes téléphoniques que nous nous proposons de décrire en partie ici.

Leur combinaison.

à elle seule à l'application de la pléthore de dispositifs de enquête, car il semble évident qu'aucune ne puisse convenir grandement à l'application aussi aux enquêtes auprès des

BRICK, J. M., et WAKSBERG, J. (1991). Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire. *Techniques d'enquête*, 17, 31-46.

BRICK, J. M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.

BRICK, J. M., WAKSBERG, J., KULP, D. et STARKER, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

CHU, A., EISENHOWER, D., HAY, M. MORGANSTEIN, D., NETTER, J. et WAKSBERG, J. (1992). Measuring the recall error in self-reported fishing and hunting activities. *Journal of Official Statistics*, 8, 19-39.

NETTER, J., et WAKSBERG, J. (1964). A study of response errors in expenditure data from household interviews. *Journal of the American Statistical Association*, 59, 18-55.

NETTER, J., et WAKSBERG, J. (1965). Response Errors in Collection of Expenditure Data from Household Interviews: An Experimental Study. (Bureau of the Census Technical Paper No. 11). Washington, DC: U.S. Government Printing Office.

WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section, American Statistical Association*, 1973, 429-434.

WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WAKSBERG, J. (1983). A note on locating a special population using random digit dialing. *Public Opinion Quarterly*, 47, 576-579.

WAKSBERG, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. (avec discussion). *Journal of Official Statistics*, 14, 119-147.

WAKSBERG, J., JUDKINS, D. et MASSEY, J. (1997). Suréchantillonnage géographique dans les enquêtes démographiques aux États-unis. *Techniques d'enquête*, 23, 69-80.

ARTICLE SOLICITÉ WAKSBERG 2001

Auteur: Gad Nathan

Gad Nathan est professeur de statistique à la Hebrew University of Jerusalem et travaille de longue date pour le Central Bureau of Statistics d'Israël. Le plus récemment à titre d'expert scientifique en chef. Il a obtenu son diplôme de doctorat au Case Institute of Technology, à Cleveland, en Ohio, et a publié de nombreux articles dans d'importantes revues statistiques, y compris *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society*, *Techniques d'enquête*, *Journal of Official Statistics* et *Sankhya*. Ses principaux domaines de recherche sont la méthodologie d'échantillonnage, l'inférence à partir d'échantillons complexes, l'interview assistée par ordinateur et les téléenquêtes. Il a occupé divers postes de professeur invité et d'expert dans plusieurs établissements universitaires et bureaux de la statistique d'Amérique du Nord et d'Europe, et a été vice-président de l'Institut international de statistique et de l'Association internationale des statisticiens d'enquête, ainsi que président de l'Israel Statistical Association et de l'Israel Public Council of Statistics.

M. Waksberg a partagé ses connaissances et son expérience avec d'autres dans de multiples endroits. Pendant de nombreuses années, il a enseigné à la Graduate School du U.S. Department of Agriculture et a été un professeur régulier du programme d'été sur les méthodes d'échantillonnage de l'Université du Michigan. Il a aussi rempli le rôle d'expert-consulte en échantillonnage et en techniques d'enquête auprès d'organismes officiels de la statistique de nombreux pays, sous le patronage de la U.S. Agency for International Development et des Nations-Unies, ainsi qu'à la demande de pays particuliers. Il a fourni des conseils aux agences statistiques de la Chine, de l'Argentine, du Brésil, du Cuba, du Venezuela, de la Turquie et du Vietnam du Sud. Il a aussi représenté les États-Unis à des colloques internationaux sur la statistique, a joué le rôle d'expert technique sous les auspices des Nations-Unies et a fait partie d'une équipe envoyée dans divers pays d'Amérique du Sud par l'*American Statistical Association* pour coordonner les activités de leur société statistique nationale. Il est membre de l'*American Statistical Association*, qui lui a conféré le titre de Fellow, de l'*Association internationale des statisticiens d'enquêtes* et de l'*Institut international de statistique*, et a participé à divers groupes de travail de la National Academy of Sciences chargés d'évaluer des programmes statistiques fédéraux particuliers. Il a été le premier lauréat du prix Roger Herring décerné par la *Washington Statistical Society*, et les sections Government Statistics et Social Statistics de l'ASA pour « innovation dans la statistique fédérale », et est un récipiendaire de la Gold Medal Award du U.S. Commerce Department. Enfin, son héritage le plus important sera peut-être le très grand nombre de collègues qui ont été inspirés par son exemple personnel, ses enseignements, son leadership, ainsi que sa gentillesse, sa bienveillance et sa compréhension.

rendre de nouveau visite à un échantillon d'unités inoccupées qui a permis de montrer qu'une proportion importante de ces unités étaient, en fait, occupées. Il a alors mis au point et appliqué une méthode de rajustement. Par la suite, après l'adoption de la loi sur le partage des revenus de 1972 qui exigeait que le Bureau produise des estimations annuelles de la population et du revenu par habitant pour les 39 000 unités gouvernementales des États-Unis, Waksberg a proposé d'utiliser des dossiers administratifs ainsi que des données d'enquêtes pour produire les estimations régionales requises de la population et du revenu par habitant. Il a entrepris des travaux de recherche sur l'appariement des dossiers de l'IRS couvrant deux années consécutives en vue d'obtenir des estimations régionales (au niveau du comté) des migrations brutes et nettes, et de la variation du niveau de revenu. Ces recherches ont mené au développement et à la mise en oeuvre d'un programme d'estimation pour petites régions qui est encore utilisé aujourd'hui.

Durant ses années à Westat depuis 1973, d'abord en qualité de statisticien principal et vice-président, et récemment comme consultant et président du conseil, Waksberg a continué de faire preuve de la même passion pour l'innovation, l'expérimentation et la qualité lorsqu'il s'efforçait de répondre aux besoins de ses clients, de mettre au point des échantillons ou d'exécuter des travaux de recherche sur les enquêtes. Lors de l'élaboration des plans de sondage de la National Health Interview Survey et de la National Health and Examination Survey du National Center for Health Statistics, il a participé à l'élaboration de méthodes innovatrices permettant de surechantillonner plus efficacement les populations minoritaires (Waksberg 1973). Les travaux qu'il a réalisés en collaboration avec Judkins et Massey fournissent des renseignements importants sur la concentration résidentielle selon la race et l'origine ethnique, qui sont essentiels à l'évaluation de l'utilité dans certaines régions géographiques du surechantillonage des populations minoritaires des personnes pauvres, une autre sous-population qu'il est souvent nécessaire de surechantillonner (Waksberg, Judkins et Massey 1997). En collaboration, il a développé la méthode Mitofsky-Waksberg d'échantillonnage à deux phases des ménages abonnés au téléphone (Waksberg 1978). Cette méthode est devenue la méthode standard d'échantillonnage par composition aléatoire (CA) aux États-Unis. En vue de l'améliorer, Waksberg a étudié le biais que la méthode de composition aléatoire introduit comparativement à l'échantillonnage basé sur une liste (Waksberg 1983; Brick et Waksberg 1991). Cette étude, qui a permis de modifier la méthode de la rendue plus efficace, a mené à une toute nouvelle méthode d'échantillonnage par CA (Brick, Waksberg, Kulp et Stater 1995). Plus récemment, il a participé à une étude sur des méthodes alternatives d'ajustement pour les ménages n'ayant pas de téléphone (Brick, Waksberg et Keeter 1996). Les travaux de Waksberg dans le domaine de

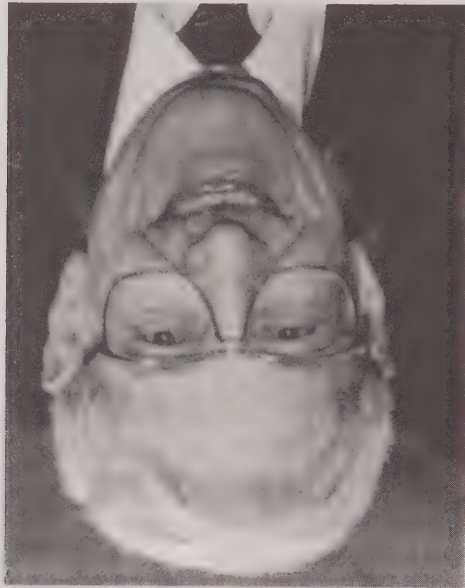
titre de commis. Il y a travaillé pendant 33 ans jusqu'au moment de sa retraite, en 1973. À ce moment-là, il était directeur adjoint du service des méthodes statistiques, de la recherche et des normes. Au début des années 60, Waksberg et Netter ont entrepris une étude classique sur l'importance de divers problèmes de mémorisation. Cet effort marquant a permis de mettre au point des méthodes visant à réduire les effets de ces problèmes grâce à une stratégie innovatrice d'échantillonnage et de collecte de données. (Netter et Waksberg 1964; Netter et Waksberg 1965). Par la suite, Joe a continué de s'intéresser à ce domaine; par exemple, il a participé à la conception et à l'analyse des résultats d'une expérience réalisée pour déterminer le sens et la grandeur des biais éventuels dans le cas d'une enquête comportant une période de rappel d'un an pour le U.S. Fish and Wildlife Service (Chu, Eisenhower, Hay, Morganstein, Netter et Waksberg 1992). Cette expérience a eu un effet important sur le remaniement de l'enquête, mais a, avant tout et par dessus tout, élargi considérablement le champ des connaissances sur le biais de réponse quand on demande aux répondants de se remémorer la fréquence de certaines activités pour diverses périodes de rappel, et indiqué des méthodes pour minimiser l'erreur quadratique moyenne du plan de telles enquêtes.

La Current Population Survey (CPS) américaine, qui est considérée aujourd'hui comme un modèle d'efficacité statistique, reflète pleinement l'influence et la contribution de Joseph Waksberg à l'époque où il était responsable de l'échantillonnage, des normes statistiques et de la recherche pour le programme des enquêtes-ménages du Census Bureau. L'amélioration des méthodes d'échantillonnage et d'estimation, y compris l'utilisation d'échantillonnage à partir de bases-listes, les méthodes d'estimation de la variance par répétition, la détermination de la taille appropriée des grappes, le traitement des événements rares et l'estimation composite sont des changements qui portent tous son empreinte. Parallèlement, Joe a joué un rôle important dans la recherche expérimentale touchant à de nouveaux scénarios de renouvellement d'échantillon et à l'utilisation d'un seul répondant par ménage, et aux effets de périodes de rappel variables sur mesure de la population active.

Aucun exposé des travaux de Joe au Census Bureau ne serait complet sans la mention de ses nombreuses contributions au programme de recensement décennal. Le programme d'évaluation qu'il a développé, conçu et dirigé pour le Recensement de 1970 en est un bon exemple. Ce programme, qui regroupait 25 projets distincts, a été qualifié à l'époque de « radical »; aujourd'hui, il sert de modèle aux programmes courants de recherche sur le recensement décennal. Durant le Recensement de 1970, lorsque les premiers résultats des travaux sur le terrain ont indiqué une surestimation grave des unités « inoccupées », Waksberg a conçu, élaboré et mis en oeuvre, en très peu de temps et avec des moyens financiers très limités, un programme innovateur d'enquête par sondage consistant à

Série Waksberg d'articles sollicités

Le comité de rédaction de *Techniques d'enquête* a décidé de publier une série d'articles annuels sollicités en l'honneur de Joseph Waksberg, pour souligner sa contribution importante à la méthodologie d'enquête. Chaque année nous inviterons un spécialiste renommé de la recherche en sondages à rédiger un article consacré à la rétrospective et à l'examen de la situation courante d'un domaine important de la méthodologie d'enquête. L'auteur reçoit un prix monétaire grâce à une subvention offerte par Westat en reconnaissance de la contribution de Joe Waksberg durant les nombreuses années où il a travaillé pour l'entreprise. L'*American Statistical Association* est chargée de la gestion financière et administrative de la subvention. L'auteur de l'article est choisi par un comité de quatre personnes désignées par *Techniques d'enquête* et l'*American Statistical Association*.



JOSEPH WAKSBERG

À l'heure actuelle, Joseph Waksberg (que tout le monde appelle « Joe ») est président du Conseil d'administration de Westat, une entreprise de recherche statistique établie à Rockville, Md. Au cours de sa carrière, qui s'étend maintenant sur plus de 60 ans, il a fait d'importantes contributions à la théorie de l'échantillonnage à développé, des applications innovatrices de la théorie et mené des travaux de recherche sur des problèmes très variés de méthodologie d'enquête. Il est l'auteur ou le coauteur de nombreux articles sur les méthodes d'échantillonnage, y compris la composition aléatoire, l'échantillonnage des populations rares, l'échantillonnage pour les enquêtes par panel et avec renouvellement de l'échantillon et le rôle de l'échantillonnage dans les recensements de population. Ses autres contributions vont de la recherche méthodologique sur l'évaluation de la population active à l'estimation régionale

et à la création de modèles pour prédire les vainqueurs la nuit des élections, en passant par l'évaluation de la qualité des données de recensement aux États-Unis, les effets du télescopage et d'autres problèmes de mémorisation sur les résultats d'enquête et l'étude des effets des récompenses monétaires sur les taux de réponse et les coûts d'enquête. Son but a été d'améliorer et la théorie et la pratique. Enfin, mais non le moindre de ses accomplissements, il a été le maître et le mentor de nombreuses générations de statisticiens. Né à Kielec, en Pologne, en septembre 1915, Joe a émigré aux États-Unis avec sa famille en 1921. En 1936, peu après avoir obtenu un diplôme en mathématiques à la City University of New York (CUNY), il s'est installé dans la région de Washington D.C. et, après un bref séjour au Navy Department, s'est joint au Census Bureau en 1940 à

Kim propose une nouvelle méthode d'estimation de la variance basée sur un modèle linéaire d'imputation par régression qui tient compte de l'imputation aléatoire. La méthode consiste à créer un ensemble de pseudovaleurs pour y , de sorte que l'estimateur classique de la variance calculé d'après ces pseudovaleurs tienne aussi compte de l'imputation. Le calcul des pseudovaleurs est décrit d'abord pour l'échantillonnage aléatoire simple, puis pour des plans de sondage complexes. L'auteur montre que la méthode est asymptotiquement équivalente à la méthode corrigée du jackknife de Rao et Sitter et étudie ses propriétés au moyen d'une étude de simulation.

Dans leur article intitulé « Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression », Raghunathan, Lepkowski, Van Hoewyk et Solenberger abordent la question importante de l'imputation dans le cas d'une structure de données complexe où il est difficile de formuler des modèles multidimensionnels complets explicites. Leur stratégie consiste à procéder à l'imputation, une variable à la fois, en la subordonnant toutefois à toutes les variables observées. Autrement dit, les imputations sont réalisées suivant une série de régressions multiples dont le modèle varie selon le type de variable visée par l'imputation.

Dans leur article, Dufour, Gagnon, Morn, Renaud et Sarnadal proposent une mesure de distance qui permet de mesurer l'incidence relative de l'ajustement de non-réponse, de la calibration et de l'interaction entre ces deux procédures. Cette mesure leur permet d'étudier et de mesurer le changement (du poids initial au poids final) qui est produit par la procédure de modification des poids. Ils utilisent cette mesure comme outil pour comparer l'efficacité de diverses méthodes d'ajustement pour la non-réponse par l'entremise d'une étude de simulation auprès des données de l'Enquête sur la dynamique du travail et du revenu. La mesure est également appliquée aux données de l'Enquête nationale longitudinale sur les enfants et les jeunes.

Ces dernières années, les efforts se sont multipliés en vue d'étudier les populations de sans-abri des grandes villes. La difficulté qu'il y a à élaborer une base de sondage et une méthode d'échantillonnage fiable et efficace, ainsi que la fluidité de la population au fil du temps rendent ce genre d'études particulièrement complexes. Le dernier article du présent numéro, publié par Ardilly et Le Blanc, décrit les méthodes d'échantillonnage et d'estimation adoptées pour une enquête menée à l'heure actuelle en France auprès des sans-abri. Y sont également décrits les problèmes et les défis que pose ce genre d'échantillonnage des personnes sans abri se fera indirectement, par le biais des services qu'elles sont susceptibles d'utiliser, comme les refuges et les services de repas. Les auteurs montrent que la méthode de partage des poids est un moyen efficace d'obtenir des poids non biaisés pour diverses périodes de référence, comme une journée moyenne ou une semaine moyenne.

Finalement, j'aimerais profiter de l'occasion pour exprimer de sincères remerciements à Frank Mayda, Gestionnaire de la production de *Techniques d'enquête*, qui a récemment pris sa retraite. L'implication qu'il a eue dans *Techniques d'enquête* depuis 1987 est inestimable. J'aimerais aussi annoncer que Eric Kanacout remplace Frank Mayda en tant que Gestionnaire de la production.

Dans ce numéro

Ce numéro de *Techniques d'enquête* contient le premier d'une série d'articles annuels sollicités publiés en l'honneur de Joseph Waksberg. Une brève description de la nouvelle série et une biographie de Joseph Waksberg précèdent l'article proprement dit. J'aimerais remercier Danny Levine d'avoir écrit la biographie de Joseph Waksberg. J'aimerais également remercier David Binder, Paul Biemer, Graham Kalton et Chris Skinner, les membres du comité de sélection, pour leur choix d'un chercheur de renom dans le domaine des enquêtes. Il est le premier auteur de la série Waksberg d'articles sollicités. Je dois des remerciements spéciaux à Graham Kalton qui, en tant que président fondateur du comité, a pris l'initiative, a négocié les arrangements nécessaires avec Westat, l'*American Statistical Association* et *Techniques d'enquêtes* pour mettre le projet en branle, et a travaillé pour rencontrer les délais prescrits par la revue pour la publication du numéro de juin.

L'auteur de l'article de 2001 de la série Waksberg est Gad Nathan. Son article, intitulé « Méthodes de téléenquêtes applicables aux enquêtes auprès des ménages – Revue et réflexions au sujet de l'avenir », fait l'historique de la méthodologie des enquêtes téléphoniques des années 1930 jusqu'à nos jours. Il évoque notamment les questions relatives au plan d'échantillonnage, aux bases de sondage, à la couverture, à la non-réponse et à la pondération. L'article se termine par l'examen de certains défis que posent les progrès technologiques les plus récents, comme le courrier électronique, Internet, le téléphone cellulaire et d'autres transformations technologiques et sociales, ainsi que des perspectives qu'ils offrent.

des estimateurs classiques de régression.

Fuller et Rao évaluent analytiquement les propriétés de l'estimation composite par régression. Ils commencent par décrire deux des premières variantes de l'estimation composite par régression, appelées estimateurs de régression modifiés (*MR1* et *MR2*), puis analysent l'efficacité et le comportement des estimations dans le temps en se servant d'un modèle simple de série chronologique pour les estimations calculées d'après les données d'enquête par panel. Ils concluent qu'un estimateur modifié, qui serait un compromis entre *MR1* et *MR2*, offrirait dans l'ensemble les propriétés les meilleures.

Dans son article, Bell compare l'application de divers estimateurs aux données de l'Australian Labour Force Survey. Il étudie notamment l'estimateur composite *AK*, la première variante de l'estimateur composite de régression appelée *MR2*, la variante de l'estimateur composite de régression de Fuller et Rao, ainsi qu'un meilleur estimateur linéaire non biaisé (*MELNB*) choisi comme étant une combinaison linéaire « optimale » des estimations fondées sur des données de panel. Il propose aussi un *MELNB* amélioré, obtenu en étalonnant le *MELNB* par rapport à des données de référence sur la population. La comparaison des estimateurs susmentionnés aux estimateurs de régression classiques porte sur l'écart entre les estimations qu'ils produisent et celles obtenues au moyen des estimateurs classiques, leurs écarts-types et leur utilité pour la désaisonnalisation et l'estimation de la tendance.

Le dernier article de la section spéciale, publié par Gambino, Kennedy et M.P. Singh, décrit l'estimateur composite de régression utilisé au Canada pour l'Enquête sur la population active. Cet estimateur se fonde sur les travaux de A. C. Singh et de ses collègues et sur le compromis proposé par Fuller et Rao. Les auteurs comparent les nouveaux estimateurs aux estimateurs de régression classiques utilisés antérieurement en les appliquant à plusieurs séries de données. Ils constatent que les nouveaux estimateurs sont ordinairement plus efficaces et plus stables et que leur utilisation pour la désaisonnalisation de la série d'estimations donne plus fréquemment de bons résultats.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 27, numéro 1, juin 2001

TABLE DES MATIÈRES

Dans ce numéro	1
Article Sollicite Waskbserg	
G. NATHAN	
Méthodes de téléenquêtes applicables aux enquêtes-ménages – Revue et réflexions sur l'avenir	7
Section spéciale sur l'estimation composite	
A.C. SINGH, B. KENNEDY et S. WU	
Estimation composite par régression pour l'Enquête sur la population active du Canada avec plan de sondage à renouvellement de panel	35
W.A. FULLER et J.N.K. RAO	
Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada	49
P. BELL	
Comparaison d'autres estimateurs pour l'Enquête sur la population active	57
J. GAMBINO, B. KENNEDY et M.P. SINGH	
Estimation composite par régression pour l'Enquête sur la population active du Canada : Evaluation et application	69
Articles Réguliers	
J.-K. KIM	
Estimation de la variance après imputation	81
T.E. RAGHUNATHAN, J.M. LEFKOWSKI, J. VAN HOEWYK et P. SOLENNBERGER	
Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression	91
J. DUFOUR, F. GAGNON, Y. MORIN, M. RENAUD et C.-E. SÄRDAL	
Mieux comprendre la transformation des poids à l'aide d'une mesure de changement	105
P. ARDILLY et D. LE BLANC	
Echantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français	117

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

- Président** G.J. Brackstone
Membres D.A. Binder
G.J.C. Hole
E. Kancourt (Gestionnaire de la production)
C. Patrick
D. Roy
M.P. Singh
R. Platek (Ancien président)

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*

Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistique Canada*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *University of Southampton, U.K.*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*

Rédacteurs adjoints

J.-F. Beaumont, P. Dick, H. Mantel et W. Yung, *Statistique Canada*
G. Nahhan, *Hebrew University, Israel*
D. Norris, *Statistique Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *The Urban Institute*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclus pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires); autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Août 2001

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2001

Publication autorisée par le ministre
responsable de Statistique Canada

JUN 2001 • VOLUME 27 • NUMÉRO 1

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

•

VOLUME 27

•

JUIN 2001

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE





SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2001

•

VOLUME 27

•

NUMBER 2





SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 2001 • VOLUME 27 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 2002

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

February 2002

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D.A. Binder
G.J.C. Hole
E. Rancourt (Production Manager)
C. Patrick

R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat, Inc.*
C. Clark, *U.S. Bureau of the Census*
J.-C. Deville, *INSEE*
J. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
D. Holt, *University of Southampton, U.K.*
G. Kalton, *Westat, Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *Joint Program in Survey Methodology*
S. Linacre, *Official National Statistics*

G. Nathan, *Hebrew University, Israel*
D. Norris, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
L.-P. Rivest, *Université Laval*
F.J. Scheuren, *National Opinion Research Center*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
E. Stasny, *Ohio State University*
R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, H. Mantel and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is CDN \$47 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et staticiens du Québec.

SURVEY METHODOLOGY
A Journal Published by Statistics Canada
Volume 27, Number 2, December 2001

CONTENTS

In This Issue	119
K. BLENK DUNCAN and E.A. STASNY Using Propensity Scores to Control Coverage Bias in Telephone Surveys	121
L.T. MARIANO and J.B. KADANE The Effect of Intensity of Effort to Reach Survey Respondents: A Toronto Smoking Survey	131
M.A. HIDIROGLOU Double Sampling	143
P. LAVALLÉE and P. CARON Estimation Using the Generalised Weight Share Method: The Case of Record Linkage	155
T. MERKOURIS Cross-sectional Estimation in Multiple-Panel Household Surveys	171
D.A. MARKER Producing Small Area Estimates From National Surveys: Methods for Minimizing use of Indirect Estimators ..	183
H. SAIGO, J. SHAO and R.R. SITTER A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data	189
D.R. BELLHOUSE and J.E. STAFFORD Local Polynomial Regression in Complex Surveys	197
D.B.N. SILVA and T.M.F. SMITH Modelling Compositional Time Series from Repeated Surveys	205
Acknowledgements	217

In This Issue

This issue of *Survey Methodology* contains papers on a variety of topics touching on coverage issues, nonresponse, imputation, survey designs, survey weighting and analysis of data from complex surveys.

In the first paper of this issue, Blenk and Stasny develop a weighting adjustment in order to reduce the coverage bias in telephone surveys while controlling the increase in variance due to weighting. The weighting adjustment is applied to *transient* households, which are households moving in and out of the telephone population during the year. It is assumed that the transient telephone population is representative of the non-telephone population. The weighting adjustment proposed is based on propensity scores for transience obtained using a logistic regression model. The proposed method and several alternatives are compared using data collected from a survey of distressed and non-distressed regions of Kentucky, Ohio, and West Virginia.

Mariano and Kadane use the information on the number of calls in a telephone survey as an indicator of how difficult an intended respondent is to reach. This permits a probabilistic division of the nonrespondents into those who will always refuse to respond and those who were not available to respond in a model of the nonresponse. It also permits an evaluation of whether the nonresponse is ignorable for inference about the dependent variable by incorporating the information on the number of calls into the model. These ideas are implemented on data from a survey in Metropolitan Toronto of attitudes toward smoking in the workplace. The results reveal that the nonresponse is not ignorable and those who do not respond are twice as likely to favor unrestricted smoking in the workplace as are those who do.

In his paper, Hidirolou unifies the nested and non-nested cases found in the double sampling theory. The nested case, also known as two-phase sampling, corresponds to the traditional case in which a first-phase sample is initially taken so that additional information may be collected. This is followed by a second-phase sample taken within the first one, which contains the variables of interest. The non-nested case reflects a situation in which both samples are selected independently from the same frame or possibly from different frames. Using the generalized difference, an estimator is proposed for both cases, and an optimal estimator that minimizes variance is developed. Variance estimation is also discussed for both cases. Numerous examples of surveys conducted at Statistics Canada illustrate the unification of both cases.

Lavallée and Caron investigate the problem of producing estimates when using record linkage methods to link two populations together. In particular, they consider the problem of producing estimates for one of the populations using a sample from the other one, assuming the two populations have been linked together. The Generalized Weight Share method is adapted to take into account the linkage weights in three different ways: (1) all links where the linkage weight is non-zero; (2) all links where the linkage weights are greater than a given threshold; and (3) the links are randomly chosen. These proposed estimators are compared with the classical approach through a simulation study.

Merkouris considers the problem of producing cross-sectional estimates with data collected from multiple panel surveys. Coverage of the cross-sectional population maybe incomplete due to individuals leaving or entering the population after the selection of the panel. By recognizing that a repeating panel survey is a special type of multiple frame survey, Merkouris is able to propose weighting strategies suitable for various multiple panel surveys. These weighting procedures can be used to combine information from the multiple panels to produce cross-sectional estimates that take into account the dynamic character of the multiple panel design.

Marker investigates survey design strategies to improve the quality of direct small area estimators, thus reducing the need for indirect, model-based estimators. Factors considered include stratification and oversampling, combining data from repeated surveys, harmonizing across different surveys, supplemental samples, and improved estimation procedures.

In their paper, Saigo, Shao and Sitter address the important problem of variance estimation under imputation for missing data. In their paper, they propose a bootstrap method that works for both smooth and non-smooth statistics, even for the case where the number of sampled clusters is small. This improves on their previously proposed bootstrap method which could suffer from serious overestimation when the number of sampled clusters is small. In addition to a bootstrap method, Saigo, Shao and Sitter also propose a repeated Balanced Repeated Replication method that captures the imputation variance in the presence of random imputation. These methods are illustrated through a simulation study.

Bellhouse and Stafford consider nonparametric local polynomial regression as an exploratory data analysis tool for data from complex surveys. They consider a single continuous regressor variable x , which is binned into a finite number of possible values, which may correspond to the precision of measurement of x , but may also be chosen otherwise. Point estimates of the local regression function, and associated variance estimates, are developed. The method is illustrated with an analysis of body mass indices from the Ontario Health Survey, and the nonparametric estimates are compared to those obtained from a parametric model.

In the final paper of this issue, Silva and Smith use a state space approach for modelling of compositional time series using data from a repeated complex survey. A compositional time series is a multivariate time series of proportions constrained to add to one at each time point. They first transform the data using an additive logistic transformation, and then model the transformed series. Estimation methods based on the Kalman filter are developed and then applied to data from the Brazilian Labour Force Survey. The Kalman filter also provides model-based estimates of variance and confidence limits for the transformed series. Estimates of trends and seasonal effects are compared to those obtained using X-11 ARIMA, and found to be generally smoother since they explicitly account for sampling errors in the raw estimates of the series.

M.P. Singh

Using Propensity Scores to Control Coverage Bias in Telephone Surveys

KRISTIN BLENK DUNCAN and ELIZABETH A. STASNY¹

ABSTRACT

Telephone surveys are a convenient and efficient method of data collection. Bias may be introduced into population estimates, however, by the exclusion of nontelephone households from these surveys. Data from the U.S. Federal Communications Commission (FCC) indicates that five and a half to six percent of American households are without phone service at any given time. The bias introduced can be significant since nontelephone households may differ from telephone households in ways that are not adequately handled by poststratification. Many households, called "transients", move in and out of the telephone population during the year, sometimes due to economic reasons or relocation. The transient telephone population may be representative of the nontelephone population in general since its members have recently been in the nontelephone population.

This paper develops a weighting adjustment for transients in an effort to reduce the bias due to noncoverage while controlling the increase in variance due to weighting. We use a logistic regression model to describe each household's propensity for transience, using data collected from a survey of distressed and non-distressed regions of Kentucky, Ohio, and West Virginia. Weight adjustments are based on the propensity scores. Estimates of the reduction in bias and the error of estimates are computed for a number of survey statistics of interest, using the propensity based weight adjustments and several alternative weight adjustments. The error in adjusted estimates is compared to the error of the standard estimate to assess the effectiveness of the adjustment.

KEY WORDS: RDD survey; Weight adjustments; Non-sampling error.

1. INTRODUCTION

The telephone is a standard mode of communication in today's world, and hence it is extremely useful for conducting surveys. Telephone surveys have come into use more and more as a growing percentage of people have phone connections. Most people who belong to the population that a survey seeks to make inferences about, the survey's target population, can be reached by phone. Therefore, the sample is drawn from the set of all people in households reachable through residential phone numbers. However, this sampling frame excludes all the people without telephone service who may compose a significant portion of some populations. It is currently estimated that in the United States, five and a half to six percent of households are without telephone service at any given time (Belinfante 2000). People without phone service tend to be different from people with service, particularly with regards to economic factors (Smith 1990). Results of the survey will not truly reflect the entire population if these differences are significant on matters of importance to the survey. The coverage bias is particularly troublesome in surveys that examine subgroups of the population with lower telephone penetration rates. These groups include people in lower income households and people who have not obtained a high school degree.

Poststratification on demographic variables associated with telephone coverage is helpful for reducing the coverage bias, but it does not completely solve the problem (Massey and Botman 1988). Another way to account for

this coverage bias is to let people who are currently without telephone service be represented by people in the survey who have not had continuous service recently. People whose phone status has changed within the last year are referred to as transients. Transients move in and out of the telephone population, possibly for economic reasons, or service interruptions during relocation. Transients who currently have phone service may be good representatives of the nontelephone population because they are included in the sampling frame, yet they have recently been part of the nontelephone population.

A weighting adjustment suggested by Brick, Waksberg and Keeter (1996) uses transients in the sample to represent the nontelephone population. They use data from the U.S. Current Population Survey (CPS) to estimate unbiased weighting class adjustments for the transient respondents in their survey. Frankel, Ezzati-Rice, Wright and Srinath (1998) also employ this weighting class adjustment, and consider two similar adjustments. Brick, Flores Cervantes, Wang and Hankins (1999) and Frankel, Srinath, Battaglia, Hoaglin, Wright and Smith (1999) evaluate these adjustments using surveys that ask questions about telephone service, but that are not subject to telephone coverage bias. These studies found that employing weight adjustments based on transient status generally led to improved estimates.

This article studies an alternative method for computing a transient weight adjustment. Our method develops a model for predicting transience using demographic variables. The weight adjustment is then based on the

¹ Kristin Blenk Duncan and Elizabeth A. Stasny, Department of Statistics, Ohio State University, Columbus, OH 43210-1247.

respondent's propensity for transience. We also compare our propensity method to the method suggested by Brick *et al.* (1996), and to a response probability method where the weight adjustment is based on the length of interruption in telephone service.

We use data from the Appalachian Poll, an RDD telephone survey conducted by the Ohio State University's Center for Survey Research during June and July of 1999. The survey was sponsored by *The Columbus Dispatch*, and compared distressed and non-distressed regions of Kentucky, Ohio, and West Virginia. The study gathered information on quality of life issues and perceptions about the Appalachian regions, and also posed a series of standard demographic questions. A stratified sample was used, and just over 400 surveys were completed from each of the six strata (Appalachian and non-Appalachian regions of Ohio, Kentucky, and West Virginia). The poll targeted English speaking adults, 18 years of age or older, residing in the three states. Coverage bias is of particular concern in this survey since telephone coverage rates are lower than usual in the distressed Appalachian regions.

In section 2, we report on the literature describing telephone and transient populations. In this section we also explore differences between these groups in our data, illustrating the concern about coverage bias. Section 2 ends with our proposed model for predicting transience. Section 3 details the various weighting procedures. In section 4 we discuss the trade-off between bias reduction and increased variance from adjusted weights, and compare the weighting schemes. The final section summarizes the findings.

2. NONTELEPHONE AND TRANSIENT TELEPHONE POPULATIONS

The target population for a telephone survey can be categorized by telephone status into four groups: continuous service households, transient households which are currently with service, transient households which are currently without service, and chronic nontelephone households. We need to know something about the size of each of these groups in order to account for coverage bias in the survey. Data from the FCC is useful for examining long term trends in the size of the nontelephone population. Not as much is known, however, about the short-term changes in phone coverage.

Keeter (1995) used panel surveys to study the dynamics of the transient phone population. In the March 1992 and 1993 CPS, it was found that 94.1% of households in the sample at both times had a phone at both time points, 2.6% at neither point, and 3.4% had a phone at one interview, but not the other. Fifty-seven percent of respondents who reported having no phone at either interview were transient. If the measurements could be taken continuously, rather than at two points in time, even more households would be labeled transient. Keeter concludes that, "a sizable minority

of nontelephone households, at the least, have recently been in the telephone population or are soon to join it. Such transient households constitute a measurable segment of telephone households and thus can provide data to characterize the nontelephone population," (Keeter 1995, page 201). The same article asserts that, "Transient telephone households are much more like nonphone households than those with continuous service," (Keeter 1995, page 209). This conclusion is based on formal tests using demographic variables from the CPS. Data from the National Survey of America's Families presented in Brick *et al.* (1999) supports Keeter's findings. Since transients make up a nontrivial proportion of the nontelephone population and transients are more similar to the nontelephone households than they are to continuous service households, it is reasonable to use data from the transients in the sample to attempt to reduce coverage bias.

In the Appalachian Poll, 140 of the 2,463 respondents, or 5.7%, replied positively to the question, "During the last twelve months has your household ever been without telephone service for one week or more?" These respondents are categorized as transients. In the Appalachian regions, the transience rate is 7.4% while the rate is only 3.9% in non-Appalachian regions.

Table 1 compares transient and nontransient households from the sample in regards to selected variables. The large differences between the two populations illustrate the need for bias reduction. People who live in transient households are much younger, have lower incomes, and they are less likely to be employed full time. They also have less access to health insurance and computers.

Table 1
Selected Characteristics of Nontransient and Transient Households

Characteristics	Nontransient	Transient
Median Age	47.0	37.5
Household income Less than \$20K	27.8%	60.0%
Employed full-time or retired	55.0%	34.5%
No health insurance	12.7%	30.0%
Owns or is buying residence	79.4%	61.4%
Computer in home	47.4%	26.4%
Not enough money for food	12.3%	42.9%

Note: Statistics are based on unweighted frequencies in the sample which oversampled from the Appalachian regions, and thus are not representative of population quantities.

A model for transience. Using the Appalachian Poll sample, we develop a logistic regression model to predict transience with demographic variables. The independent variables used to predict transience are age, employment status, race, income, and region. The model is described in the Appendix. Education and tenure are also good predictors of transience, but they are strongly correlated with

the other variables in our model, and thus, we chose not to include them. For a comparison of models that predict telephone coverage, see Smith (1990). We will use our model in the propensity weighting adjustment described in the following section.

3. WEIGHT ADJUSTMENTS

We consider several weighting schemes that attempt to account for the coverage bias inherent in telephone surveys. Each of these schemes is compared to the actual weighting procedure used for the Appalachian Poll. In the standard procedure, a base weight was calculated for each respondent. This adjustment is $(\# \text{ adults in household}) / (\# \text{ voice telephone lines})$, or the inverse of the respondent's probability of being in the sample. Then weights were raked in each of the six strata to agree with 1990 Census proportions for age group, education level, and gender. Finally, the weights were scaled to the sample sizes within the six strata.

3.1 Length of Disconnect

Respondents to the Appalachian Poll who replied "yes" to the question about an interruption in phone service of one week or longer were then asked how many days they were without service in the last year. A simple approach to the coverage bias problem is to give transients a weight adjustment inversely proportional to the fraction of the year that they were with service. For example, a person who has only had service for six months out of the last twelve receives a weight of two, thus representing himself and one other person in the population with a six-month disconnect who is currently without service.

This naïve approach is included in the analysis for comparison with other schemes. It is referred to as the day scheme (DAY). Weight adjustments are calculated as $365/(365 - \# \text{ days without service})$. This weight adjustment is applied after the base weight described above, and before the weights are raked.

While this approach is logical, it is not practical for controlling variance. It is usually considered undesirable to use weighting factors larger than three. In fact, for many large surveys conducted by the U.S. Census Bureau, if weighting factors are larger than two, respondents are merged into larger groups and a group weight is calculated in order to obtain lower weighting-adjustment factors; see, for example, CPS (1978).

This simple approach becomes more practical when respondents are grouped by the length of their interruption in service. In a scheme called day group (DAYG), transients are grouped into quartiles across the entire sample by length of interruption in phone service. These quartiles correspond to interruptions of one week, more than one week but less than three weeks, three weeks to two months, and more than two months. The weight adjustment for each group is $365/(365 - \text{avg. } \# \text{ days without service})$, and it is also applied after the base weight, prior to raking. This

grouping procedure is helpful for reducing the variance caused by extremely long interruptions.

3.2 Weighting Class Adjustment Scheme

Brick *et al.* (1996) also implement a response probability adjustment to reduce coverage bias. Under their procedure, they partition the target population into the four components described in section 2: t_1 is the number of persons living in continuous service households; t_2 is the number of persons living in transient households that currently have service; t_3 is the number of persons living in nontelephone households that have not had any service in the last year; and t_4 is the number of persons living in transient households that are currently without service. The response probability model the authors use assumes that $t_3 = 0$. With this assumption, an unbiased weight adjustment is $A = (t_2 + t_4)/t_2 = 1 + (t_4/t_2)$, the inverse of the proportion of the transient population that currently has service. Unfortunately, these population quantities are unknown and must be estimated. Following the lead of Brick *et al.*, we use CPS data to estimate $t_1 + t_2$, the number of persons who currently have service, and t_4 ; call these estimates $\hat{t}_1 + \hat{t}_2$ and \hat{t}_4 , respectively. From the Appalachian Poll, separate estimates of t_1 and t_2 are available; designate these estimates as t_1^* and t_2^* , respectively. Since the estimates come from different surveys, ratios are used in the weight adjustment, and A is estimated by

$$A' = 1 + \frac{\frac{\hat{t}_4}{\hat{t}_1 + \hat{t}_2}}{\frac{t_1^*}{t_1^* + t_2^*}}. \quad (1)$$

Some persons are more likely to live in nontelephone households than others, so Brick *et al.* classified transients into cells based on characteristics associated with not having a telephone, and computed the weight adjustment for each cell. Four classification schemes, which categorized respondents by either education or tenure, length of interruption, and race/ethnicity were considered.

Brick *et al.* found schemes that classified respondents as transients if they had an interruption of one week or more to be superior to schemes that used a cut-off of one month, so for the Appalachian Poll data we use the one-week cut-off. Due to the small number of Hispanics in the Appalachian Poll sample, we do not categorize by ethnicity. Thus, for our analyses, the cell classifications for two schemes that use the method described by Brick *et al.* (1996) are defined as follows:

BWKE – households that had a service interruption of one week or more within categories defined by education (less than high school, high school diploma, college diploma or above) and race (black, non-black); and

BWKT – households that had a service interruption of one week or more within categories defined by tenure (own/other, rent) and race.

The disadvantage of using these schemes in our study is that the estimates needed from the CPS are available by state, but not by region since the CPS does not sample from all counties. Persons in Appalachian regions are less likely to have telephones, but we cannot account for this with the available CPS data. Even when we consider statewide data, the sample size of the CPS is not large enough to get reliable values of \hat{t}_4 in all of the cells. For example, in 1999 the CPS did not sample any blacks with a college degree or higher who live in Kentucky and do not have telephone service. Thus, the weighting cell adjustments computed for use with the Appalachian Poll are based on CPS data from the three states combined.

3.3 Raking Ratio Adjustment

Lohr (1999) explains the use of raking ratio estimates to adjust for nonresponse in surveys. We propose a similar use of raking to account for coverage bias. We estimate the proportion of the population with continuous telephone service, and then use raking to allow transients in the sample to represent the portion of the population without continuous telephone service.

The percent of households without continuous service is estimated by

$$1 - \left(\frac{\tilde{t}_1 + \tilde{t}_2}{\tilde{t}_1 + \tilde{t}_2 + \tilde{t}_4} \right) \left(\frac{t_1^*}{t_1^* + t_2^*} \right), \quad (2)$$

where $\tilde{t}_i, i = 1, 2, 4$, is obtained from the FCC data. The first fraction estimates the proportion of households that currently have service, and the second fraction estimates the

proportion of nontransient households among households with service. Again, we assume that $t_3 = 0$. The FCC gives telephone penetration rates by state, but not by region. Data from the 1990 Census does give penetration rates by county, but rates changed from 1990 to 1999. Therefore, to estimate the 1999 regional penetration rate, we maintained a constant ratio of percent of households without a phone in the non-Appalachian regions to percent of households without a phone in the Appalachian regions and adjusted the 1990 Census regional rates to match the 1999 state rates. Table 2 gives the data we used to compute the 1999 state rates, and the resulting estimates.

In a scheme referred to as transient raking, or TRAK, transient status is included as a control variable for raking along with age, gender, and education level. The totals we used for raking by transient status are given in Table 2.

3.4 A New Propensity Weighting

An estimated propensity score is sometimes used to create a weight adjustment to account for nonresponse in surveys where some variables are known for the nonrespondents. For example, in a face-to-face household interview the interviewer knows the address of the nonrespondent and may have information about the person's race, gender, and age. A logistic regression model that describes propensity for response is developed, and respondents are assigned a weight of $1/\hat{p}$, where \hat{p} is the estimated propensity to respond (Little and Rubin 1987). This procedure gives higher weights to sampled households that are more similar to the nonrespondents. Since there is typically no data on the excluded nontelephone population in telephone surveys, a modified approach is taken to using a propensity score. We only adjust the weights for the transients since they will represent the missing part of the sample: weights for nontransients remain unadjusted. The

Table 2
Computation of Transient Status Raking Totals

	Kentucky		Ohio		West Virginia	
	Ap	Non-Ap	Ap	Non-Ap	Ap	Non-Ap
Appalachian Poll Data						
Sample Size	412	407	413	405	411	415
# transients in sample	38	19	18	13	36	16
Percent of sample without cont. service	9.2	4.7	4.4	3.2	8.8	3.9
Census and FCC Data						
1990 State % no phone	10.2	10.2	4.7	4.7	10.3	10.3
1990 Region % no phone	19.1	8.2	11.7	4.5	14.3	8.4
1999 State % no phone	6.7	6.7	5.2	5.2	7.3	7.3
Percent of state pop. living in region	18.6	81.4	2.6	97.4	31.8	68.2
Estimates						
Ratio of Non-Ap to Ap noncoverage	0.429	0.429	0.385	0.385	0.587	0.587
Estimated 1999 region % no phone	12.5	5.4	13.0	5.0	10.1	6.0
Estimated % of pop. without cont. service	20.6	9.8	16.7	8.1	18.0	9.6
Desired # of transients in sample	85	40	69	33	74	40

weight adjustment for transients is $1/(1 - \hat{p})$, where \hat{p} , the estimated propensity for transience, is described by the model in section 2.1. Households with a higher estimated propensity for transience may be more representative of the nontelephone population and they receive higher weight adjustments. This adjustment is applied to the base weight, and the scheme is called propensity (PROP).

Transience is not that common, and most estimated propensity scores are fairly low. In the PROP scheme, the average weight adjustment for a transient household is 1.167. This adjustment is not large enough for transients to represent themselves and the entire nontelephone population. That is, when the weights are scaled to sum to the population size, the sum of the final weights for transients is less than the size of the transient population. To account for this under-representation, the propensity weight adjustment is applied, and then transient is used as a control variable for raking along with age, education, and gender. The estimated population sizes for transients are computed as in section 3.3. This weighting scheme is called augmented propensity, or AUGP.

4. FINDINGS

The analysis and comparison of the adjustment schemes presented here parallels the analysis performed by Brick *et al.* (1996). We first discuss the change in variance resulting from adjusting the weights to reduce coverage bias and present a statistic for measuring the relative variability. Then, the schemes are evaluated by comparing the variance of adjusted estimates to the mean squared error of the standard estimate.

4.1 Changes in Variability

The goal of the adjustment schemes is to decrease coverage bias while controlling variance. Adjustment of the weights to reduce the bias increases the variability of the weights, hence increasing the variance of the estimates. Kish (1992) gives a formula for measuring the increase in

variance due to unequal weights. Brick *et al.* (1996) refer to this expression as the variance inflation factor (VIF). The VIF can be written as

VIF = 1 + [CV(weights)]², (3)

where CV(weights) is the coefficient of variation of the weights. A VIF ratio is computed to compare the VIF of a new weighting scheme to that of the standard weighting scheme. Table 3 gives VIF ratios for the six strata in the Appalachian Poll data under each scheme described in section 3. A VIF ratio of 1.12, for example, indicates an increase in variance of 12 percent over the variance using the standard weighting scheme. The VIF ratio values are reasonable for all schemes except the DAY scheme which sees an average variance increase of 300 percent. The VIF ratio values for our PROP scheme are all very close to one, suggesting that the PROP weight adjustments will not increase the variance of our estimates.

4.2 Coverage Bias Reduction

Estimates of seventeen population proportions using survey variables from the Appalachian Poll were calculated for the standard weighting procedure and for each of the seven adjustment schemes (see Table 4 for a list of the seventeen variables). WesVar software was used to calculate standard errors for these estimates by means of replication. We would like to assess the effectiveness of each scheme for reducing the coverage bias on these seventeen characteristics. Estimates from an independent source that are free of telephone coverage bias would be ideal for such an assessment. Unfortunately, such benchmarks are unavailable and some model assumptions are necessary in order to perform an evaluation. We assume that the weight adjustment procedures reduce the coverage bias. Thus the difference between the standard estimate and the adjusted estimate is considered to be an unbiased estimate of the decrease in coverage bias resulting from the adjustment. The assumption favors the adjusted estimates, considering them to be unbiased.

Table 3
Ratios of Variance Inflation Factor Due to Weight Adjustment

Region	Ratio of scheme's VIF to standard weight's VIF						
	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Non-Appalachian Ohio	0.999	0.997	1.004	1.023	1.063	0.999	1.061
Appalachian Ohio	1.480	1.016	1.039	1.091	1.331	0.999	1.336
Non-Appalachian Kentucky	4.151	1.040	1.018	1.054	1.030	0.999	1.029
Appalachian Kentucky	2.433	1.069	1.045	1.042	1.129	1.003	1.145
Non-Appalachian West Virginia	6.331	1.027	1.010	1.029	1.020	0.999	1.024
Appalachian West Virginia	2.935	1.085	1.058	1.053	1.116	1.005	1.119
Scheme Average	3.055	1.039	1.029	1.049	1.115	1.001	1.119

Table 4
Estimated Reduction in Bias and Bias Ratio for Selected Characteristics

Characteristic	Standard estimate		Estimated reduction in bias								Bias Ratio					
	Estimate	St. error	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Owens Home																
Non-Appalachian Ohio	72.2	3.1	0.6	0.5	0.5	1.2	1.4	0.1	1.6	0.2	0.2	0.2	0.4	0.5	0.0	0.5
Appalachian Ohio	75.4	2.8	4.4	0.6	0.6	2.1	3.2	0.3	3.5	1.6	0.2	0.2	0.8	1.1	0.1	1.2
Non-Appalachian Kentucky	68.6	3.1	7.2	0.8	0.9	1.8	1.5	0.2	1.5	2.3	0.3	0.3	0.6	0.5	0.1	0.5
Appalachian Kentucky	80.5	2.2	2.9	0.8	0.3	1.3	0.3	0.0	0.3	1.3	0.3	0.1	0.6	0.1	0.0	0.1
Non-Appalachian West Virginia	80.0	2.3	14.2	1.6	0.9	1.9	1.4	0.2	1.4	6.1	0.7	0.4	0.8	0.6	0.1	0.6
Appalachian West Virginia	81.9	2.2	8.2	0.7	-0.4	0.5	-0.3	0.0	-0.2	3.7	0.3	-0.2	0.2	-0.1	0.0	-0.1
No Health Insurance																
Non-Appalachian Ohio	7.3	1.7	0.0	-0.1	-0.6	-1.4	-1.7	-0.1	-1.8	0.0	-0.1	-0.4	-0.8	-1.0	-0.1	-1.1
Appalachian Ohio	12.6	2.1	0.9	0.1	0.3	0.3	0.5	0.1	0.6	0.4	0.1	0.1	0.2	0.3	0.0	0.3
Non-Appalachian Kentucky	8.8	1.8	1.8	0.4	0.2	0.3	0.0	0.1	0.1	1.0	0.2	0.1	0.2	0.0	0.0	0.0
Appalachian Kentucky	22.2	2.4	3.4	0.1	-0.1	-0.2	-0.8	-0.4	-1.5	1.4	0.0	0.0	-0.1	-0.3	-0.2	-0.6
Non-Appalachian West Virginia	14.2	2.1	-4.8	-0.5	-0.7	-1.0	-1.2	-0.3	-1.4	-2.3	-0.2	-0.3	-0.5	-0.6	-0.1	-0.7
Appalachian West Virginia	24.6	2.5	2.5	-0.8	-1.7	-1.3	-2.7	-0.6	-3.0	1.0	-0.3	-0.7	-0.5	-1.1	-0.2	-1.2
Not enough Money for Food																
Non-Appalachian Ohio	10.8	1.9	-0.7	-0.6	-0.9	-1.6	-2.2	-0.1	-2.1	-0.4	-0.3	-0.5	-0.9	-1.2	0.0	-1.2
Appalachian Ohio	16.2	2.5	-4.7	-0.8	-0.6	-1.3	-3.3	-0.2	-3.4	-1.9	-0.3	-0.3	-0.5	-1.3	-0.1	-1.4
Non-Appalachian Kentucky	11.4	2.4	-3.3	-0.8	-1.3	-1.7	-1.6	-0.4	-1.8	-1.4	-0.3	-0.5	-0.7	-0.7	-0.2	-0.8
Appalachian Kentucky	20.2	2.4	-7.4	-2.3	-2.1	-2.1	-3.8	-0.4	-3.8	-3.1	-1.0	-0.9	-0.9	-1.6	-0.2	-1.6
Non-Appalachian West Virginia	14.0	2.1	4.3	-0.1	-1.0	-1.4	-1.7	-0.3	-1.8	2.1	0.0	-0.5	-0.7	-0.8	-0.2	-0.9
Appalachian West Virginia	16.4	2.0	1.5	-0.7	-1.0	-0.9	-2.2	-0.5	-2.6	0.8	-0.3	-0.5	-0.4	-1.1	-0.3	-1.3
Computer in Home																
Non-Appalachian Ohio	60.1	3.0	0.4	0.3	0.6	1.2	1.3	0.1	1.4	0.1	0.1	0.2	0.4	0.5	0.0	0.5
Appalachian Ohio	40.0	3.0	1.2	0.2	0.3	0.8	1.8	0.1	2.0	0.4	0.1	0.1	0.3	0.6	0.0	0.7
Non-Appalachian Kentucky	44.5	3.0	6.7	0.9	0.8	1.1	0.9	0.2	1.0	2.3	0.3	0.3	0.4	0.3	0.1	0.3
Appalachian Kentucky	29.7	2.3	1.9	1.0	0.9	1.1	2.3	0.0	1.9	0.8	0.4	0.4	0.5	1.0	0.0	0.8
Non-Appalachian West Virginia	46.2	2.6	7.6	0.6	1.1	1.2	1.5	0.3	1.6	2.9	0.2	0.4	0.4	0.6	0.1	0.6
Appalachian West Virginia	36.1	2.7	4.3	1.0	0.3	0.4	0.2	0.3	0.5	1.6	0.4	0.1	0.2	0.1	0.1	0.2
Summary of Seventeen Variables																
Mean absolute value			0.032	0.005	0.006	0.009	0.013	0.002	0.014	1.396	0.235	0.620	0.412	0.885	0.075	0.885
Median absolute value			0.022	0.005	0.006	0.011	0.014	0.001	0.014	0.995	0.240	0.245	0.420	0.605	0.055	0.665

Note: In addition to the four proportions listed in the table, the summary of seventeen variables includes worry about income, better off economically in the 1990's, dissatisfied with own net worth, married, have children, unemployed, college graduate, in good or excellent health, serious illness in household, no family doctor, satisfied with own housing, very safe drinking water, and internet access in home.

Using our assumption, we compare the estimate from each scheme to the standard estimate. The reduction in coverage bias is estimated by the difference between the standard estimate and the adjusted estimate. There are seven different estimates of the bias reduction, one for each scheme. The estimated reduction in bias is given by

$$b_i = \hat{p}_s - \hat{p}_i,$$

(4)

where b_i is the estimated bias reduction using scheme i , \hat{p}_s is the standard estimate, and \hat{p}_i is the estimate from adjustment scheme i . Estimated reductions in bias for four

characteristics by the six strata are given in Table 4 for each scheme. For the characteristics owns home, not enough money for food, and computer in home, the direction of the bias is fairly consistent across schemes and regions. Reassuringly, the bias is in the expected direction for these characteristics, with fewer people owning homes, more people not having enough money for food, and fewer people having computers in their homes, than is indicated by the estimates using the standard weighting scheme. For health insurance, the direction of the bias is mostly consistent across regions. The standard estimate is biased upward for Appalachian Ohio and non-Appalachian

Kentucky, and generally biased downward in the other regions.

The absolute size of the reduction in bias by itself is not fully meaningful, because it does not account for the amount of sampling error associated with the estimate. Therefore, we also calculate the bias ratio, as in Brick *et al.* (1996). The bias ratio for scheme i , r_i , is given by

$$r_i = \frac{b_i}{\text{se}(\hat{p}_s)}, \quad (5)$$

where $\text{se}(\hat{p}_s)$ is the standard error of the standard estimate. Table 4 also gives the bias ratio for the selected estimates. DAY, TRAK, and AUGP give the largest bias ratios; for these adjustment schemes the bias is not negligible when we consider the standard error. DAYG and PROP have low bias ratios, indicating that the bias reduction is small compared to the error of the estimate.

4.3 Mean Square Error

Since the standard estimates are thought to be biased, error should be measured with mean square error rather than variance. The MSE of the standard estimate is approximated by

$$\text{mse}_i = \text{var}(\hat{p}_s) + b_i^2 \quad (6)$$

for each adjustment scheme. Recall that we are assuming the adjusted estimates are unbiased, so that the mean square errors of these estimates are equal to their variances. The variance of the adjusted estimates can be approximated by two methods. The first approximation is obtained by multiplying the VIF ratio in Table 3 by the variance of the standard estimate. Alternatively, we can use the variance of the adjusted estimate obtained from replication methods.

The error of the adjusted estimate is compared to the error of the standard estimate in the mean square ratio (MSR). Using the VIF variance, the estimated MSR is given by

$$\text{msr}_{\text{VIF}_i}(\hat{p}) = \frac{100 \times \text{VIF Ratio}_i \times \text{var}(\hat{p}_s)}{\text{mse}_i(\hat{p})}. \quad (7a)$$

For the replication variance, the estimated MSR is given by

$$\text{msr}_{\text{VAR}_i}(\hat{p}) = \frac{100 \times \text{var}_i(\hat{p})}{\text{mse}_i(\hat{p})}, \quad (7b)$$

where $\text{var}_i(\hat{p})$ is the estimated variance of the adjusted estimate, obtained through replication. An MSR of 100 indicates that the variance of the adjusted estimate is exactly equal to the mean squared error of the standard estimate. An MSR above 100 means the variance of the adjusted estimate is larger than the MSE of the standard estimate, and the bias/variance trade-off for the scheme is not favorable. An MSR below 100 means that the adjusted estimate is an improvement over the standard estimate in terms of overall error.

Table 5 gives estimated MSR values for selected survey variables from the Appalachian Poll, and a summary of these values for seventeen variables from each adjustment scheme. The MSR estimates vary between regions and between schemes. The msr values computed using the two different variances also differ, but the summary values are similar for both variances. The DAY scheme has the highest msr values, indicating that this weight adjustment is not worthwhile because it increases the variance too much. TRAK and AUGP have the lowest mean and median msr values, though these schemes produced unfavorable estimates for a few characteristics as indicated by the high maximum msr values. The weighting class adjustment schemes BWKE and BWKT performed well and their maximum estimated mean square ratio values are fairly low. All of the msr values for the PROP scheme are near 100, suggesting that the overall error in estimates computed with this scheme is comparable to the error in the standard estimates.

5. CONCLUSIONS

While telephone use is commonplace, telephone surveys will always contain some bias since nontelephone households are excluded from the sampling frame, and the non-telephone population has characteristics that differ from those of the telephone population. Coverage bias is alleviated by poststratification on variables such as income and education and may not be a problem in some instances. However, for surveys that target poor or rural areas where telephone penetration rates are lower, the coverage bias is a large concern.

We have proposed a few new methods for reducing the coverage bias by adjusting the weights of respondents in the transient population. We compared the resulting estimates to those from other existing methods. In the analysis of these methods, it was assumed that the adjusted estimates are unbiased. In the absence of unbiased benchmark estimates this assumption cannot be validated. The mean square ratios presented here are likely to be biased downward since the bias of the adjusted estimate is not included. The estimated MSR is still useful for comparing methods, however, and gives a good measure of the effectiveness of the weight adjustments.

As anticipated, the DAY method was found to have too much variability to be useful. The day group (DAYG) method appears to perform better, but most of the mean square ratios for this scheme are close to 100, meaning that we do not see a large improvement over the standard estimate. The advantage of this scheme lies in its simplicity. The weight adjustment is easy to apply and does not require auxiliary data.

Table 5
Mean Square Ratio for Selected Characteristics

Characteristic	VIF Mean Square Ratio							Variance Mean Square Ratio						
	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Owens Home														
Non-Appalachian Ohio	96.1	97.2	97.3	88.1	87.5	99.8	84.5	98.6	98.2	98.2	88.4	81.8	99.9	78.7
Appalachian Ohio	42.3	97.4	98.7	68.9	57.9	99.1	52.5	71.7	96.5	89.6	71.1	51.6	99.2	46.6
Non-Appalachian Kentucky	63.9	97.6	94.5	77.7	83.1	99.4	83.5	21.9	98.2	92.8	75.7	81.3	98.8	80.1
Appalachian Kentucky	89.3	96.1	102.3	77.0	110.6	100.3	112.4	116.0	100.7	104.1	88.4	119.2	100.0	118.3
Non-Appalachian West Virginia	16.6	71.2	89.1	62.3	75.8	99.2	75.5	28.6	81.1	94.4	71.4	85.5	100.1	84.2
Appalachian West Virginia	20.2	98.4	103.2	100.3	109.3	100.5	110.6	43.5	106.0	101.1	104.8	108.8	99.1	108.9
No Health Insurance														
Non-Appalachian Ohio	99.9	99.0	88.2	61.4	51.9	99.5	48.5	98.8	100.5	112.1	101.7	82.0	101.9	76.1
Appalachian Ohio	126.8	101.3	102.1	106.5	125.1	99.9	123.5	92.3	98.8	95.6	94.0	105.8	99.0	100.4
Non-Appalachian Kentucky	206.4	99.9	100.9	102.8	103.0	99.8	102.7	39.0	87.9	90.7	86.2	97.8	96.6	95.5
Appalachian Kentucky	82.7	106.7	104.4	103.7	102.1	97.9	84.3	53.5	109.9	104.9	105.3	114.1	100.0	100.5
Non-Appalachian West Virginia	100.2	97.1	90.6	84.0	77.7	97.9	71.3	136.6	99.2	94.0	89.7	90.6	100.8	84.3
Appalachian West Virginia	149.6	99.2	74.1	83.5	52.8	95.7	46.7	107.0	96.5	75.1	84.3	51.5	96.7	45.4
Not enough Money for Food														
Non-Appalachian Ohio	86.5	90.5	80.5	57.9	45.2	99.7	45.6	105.2	100.8	104.3	94.5	66.3	102.0	67.0
Appalachian Ohio	31.9	92.9	97.4	86.1	48.6	99.4	46.4	68.5	98.2	96.8	90.2	69.4	101.1	66.4
Non-Appalachian Kentucky	139.1	94.1	78.2	69.5	69.5	96.7	64.3	320.7	96.8	91.9	85.7	77.6	100.1	68.5
Appalachian Kentucky	22.3	55.8	57.6	58.7	31.0	97.2	31.9	30.5	68.4	68.5	69.5	36.8	100.2	38.5
Non-Appalachian West Virginia	117.3	102.6	82.4	71.0	59.6	97.7	57.2	105.7	101.9	94.7	88.3	71.5	101.6	68.5
Appalachian West Virginia	181.6	97.0	84.1	88.5	50.4	94.1	39.9	92.2	98.8	89.6	92.9	59.0	97.5	48.3
Computer in Home														
Non-Appalachian Ohio	98.1	98.5	96.4	88.2	88.1	99.8	86.1	99.5	99.5	102.0	102.1	106.3	100.6	102.8
Appalachian Ohio	127.2	101.2	103.1	101.2	96.2	99.7	92.5	116.0	99.6	101.2	96.5	94.1	99.1	86.5
Non-Appalachian Kentucky	67.7	94.9	94.4	92.7	93.6	99.5	92.8	27.1	93.7	91.5	89.7	90.3	98.2	88.4
Appalachian Kentucky	147.1	89.0	91.7	85.1	55.7	100.3	68.4	58.9	81.1	85.5	79.5	46.8	100.9	66.6
Non-Appalachian West Virginia	66.8	96.9	86.6	85.8	76.1	98.5	73.5	59.6	95.8	85.1	85.3	72.9	98.6	68.6
Appalachian West Virginia	82.7	95.6	104.4	103.0	111.2	99.6	108.2	41.8	88.1	101.6	99.9	113.3	98.3	107.0
Summary of Seventeen Variables														
Mean	137.6	97.5	94.3	92.2	85.2	99.3	83.8	125.2	99.1	97.1	96.8	96.0	100.2	93.5
Median	107.5	99.0	99.1	97.1	89.8	99.8	86.3	94.8	98.9	98.7	98.5	98.0	100.0	92.4
Minimum	10.9	55.8	0.9	57.9	4.1	94.1	5.7	7.0	68.4	43.1	62.1	7.6	94.6	6.0
Maximum	607.7	108.5	104.8	109.1	133.1	100.5	133.5	695.2	140.8	144.5	147.5	593.8	116.7	545.4
Percent below 100	47.1	60.8	61.8	58.8	65.7	87.3	67.6	63.7	62.7	56.9	58.8	53.9	58.8	58.8

Note: In addition to the four proportions listed in the table, the summary of seventeen variables includes worry about income, better off economically in the 1990's, dissatisfied with own net worth, married, have children, unemployed, college graduate, in good or excellent health, serious illness in household, no family doctor, satisfied with own housing, very safe drinking water, and internet access in home.

The weighting class adjustment schemes have the benefit of giving more weight to respondents in cells where the likelihood of having a phone is lower. For these schemes, greater bias reduction was seen in variables correlated with the classification variables. For example, home ownership and computer ownership are positively correlated, and the BWKT scheme, which classified respondents by home ownership, produced estimates of the percent of households with a home computer that were consistently lower than the standard estimates. Table 5 shows that the BWKE and BWKT schemes produce an improved estimate most of the time. It should also be noted that when these schemes produce an estimate that it not an improvement, the increase in variance remains fairly small. The weighting class adjustment method works well for samples of large populations, such as states or countries, since the outside data needed to compute the adjustments is readily available. The method

is more difficult to use for very specific samples such as counties.

The raking ratio adjustment, TRAK, produced a number of very favorable estimated MSR values. With this scheme we were able to account for the difference in telephone penetration rates by region, but not the differences across other demographic characteristics. Variability was introduced when we estimated the regional rates from the state rates, thus, as with the weighting class adjustment, the scheme works better for samples of larger populations. While the mean and median estimated MSR values were low for this scheme, the scheme also produced some high mean square ratios. The higher ratios occurred in Ohio where the percent of transients in the sample was low compared to the estimated percent without continuous service.

The propensity adjustment alone, PROP, provided too little reduction in bias to be worthwhile. The propensity adjustment is advantageous, however, because it allows us to account for differences in the likelihood of having telephone service without using outside data. When used in conjunction with raking, the propensity based scheme AUGP produced good results.

There are many issues to consider when determining which adjustment scheme is preferred. As mentioned previously, the weighting class adjustment schemes BWKE and BWKT are difficult to implement if you have a very specific target population. These schemes are fairly conservative, however, in that they typically reduce the bias without increasing the variance. The schemes that employed raking usually performed better than the weighting class adjustment schemes, but the larger weight adjustments sometimes led to increased variances. It may be advisable to compute estimates using several schemes and then determine which scheme offers the best bias-variance trade-off.

Brick *et al.* (1996) note that these weight adjustments for telephone coverage should be more beneficial in reducing mean squared error when the sample size of the survey is large. As the sample size increases, the bias ratio increases since the bias is unaffected but the standard error of the estimate, which is in the denominator, decreases.

The findings suggested by this study and others indicate that the adjustments could be useful for many estimates from telephone surveys and should be seriously considered. The benefits of adjustment appear to outweigh the penalties in the weighting class adjustment schemes, the raking scheme, and the augmented propensity scheme. In light of the smaller sample size and special target population of the Appalachian Poll, generalizations of these findings should not be made until the methods receive further evaluation. These weight adjustments still need to be tested using a survey that is free of coverage bias, one that includes nontelephone households in the sampling frame and collects information on telephone status, in order to assess the validity of the assumptions. Data from the National Survey of America's Families, or the National Health Interview Survey may be appropriate for evaluating the adjustment methods and the assumptions.

ACKNOWLEDGEMENTS

This work was supported in part by a fellowship from the Center for Survey Research at the Ohio State University. We thank Dr. Paul Lavrakas and the Center for Survey Research for allowing us to use the Appalachian Poll data. We would also like to thank the referees for their helpful comments.

APPENDIX

Logistic Regression of Transient Status

Below is our model for predicting transient status. Most of the variables in the model relate to socioeconomic status. The coefficients indicate that young people, those with low income, those who are not employed full-time, American Indians and African Americans, and residents of distressed counties have higher propensities for transience. The high significance level of the Hosmer and Lemeshow test indicates a very good fit of the model. The large area under the ROC curve tells us that the model discriminates well.

Variable Coding

Age

- 0 - "Refused" (Count = 9)
- 1 - 18 to 29 years
- 2 - 30 to 44 years
- 3 - 45 to 59 years
- 4 - over 60

Low Income

- 0 - Household income over \$20,000 or refused
- 1 - Household income under \$20,000

Employment Status

- 0 - Employed full-time or retired
- 1 - Other (refused, part-time, housekeeper, student, unemployed, other)

Race

- 0 - Caucasian, Alaskan Native, Hispanic, or Asian
- 1 - American Indian, African-American, Black, or other

Appalachian

- 0 - Does not live in a distressed county of KY, OH, or WV
- 1 - Lives in a distressed county

Kentucky/West Virginia

- 0 - Ohio
- 1 - Kentucky or West Virginia

Results

Variables in the Equation

Variable	B	S.E.
Age (Refused)	-2.107	12.160
Age (18-29)	2.006	0.357
Age (30-44)	1.664	0.347
Age (45-59)	1.064	0.364
Low Income	1.358	0.189
Employment Status	0.397	0.187
Race	1.136	0.292
Appalachian	0.531	0.196
KY/WV	0.567	0.216
Constant	-5.712	0.401

Hosmer and Lemeshow Goodness of Fit Test

Chi-Square	3.568
Degrees of Freedom	8
p-value	0.894

ROC Curve

Area under the Curve	0.782
----------------------	-------

REFERENCES

- BELINFANTE, A. (2000). Telephone Subscribership in the United States. Industry Analysis Division, Common Carrier Bureau, Federal Communications Commission, Washington, D.C. 20554.
- BRICK, J.M., FLORES CERVANTES, I., WANG, K. and HANKINS, T. (1999). Evaluation of the use of data on interruptions in telephone service. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 376-381.
- BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.
- CURRENT POPULATION SURVEY (1978). Current Population Survey: Design and Methodology. Technical Paper 40. Department of Commerce, Bureau of the Census, Washington, D.C.
- FRANKEL, M.R., EZZATI-RICE, T., WRIGHT, R.A. and SRINATH, K.P. (1998). Use of data in interruptions in telephone service for noncoverage adjustment. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 290-295.
- FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. and SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 934-939.
- KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.
- KISH, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200.
- LITTLE, R., and RUBIN, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons, 55-60.
- LOHR, S. (1999). *Sampling: Design and Analysis*. New York: Duxbury Press, 255-287.
- MASSEY, J., and BOTMAN, S. (1988). Weighting adjustments for random digit dialed surveys. In *Telephone Survey Methodology*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls and J. Waksberg). New York: John Wiley and Sons, 143-160.
- SMITH, T. (1990). Phone home? An analysis of household telephone ownership. *International Journal of Public Opinion Research*, 2, 369-390.

The Effect of Intensity of Effort to Reach Survey Respondents: A Toronto Smoking Survey

LOUIS T. MARIANO and JOSEPH B. KADANE¹

ABSTRACT

The number of calls in a telephone survey is used as an indicator of how difficult an intended respondent is to reach. This permits a probabilistic division of the non-respondents into non-susceptibles (those who will always refuse to respond), and the susceptible non-respondents (those who were not available to respond) in a model of the non-response. Further, it permits stochastic estimation of the views of the latter group and an evaluation of whether the non-response is ignorable for inference about the dependent variable. These ideas are implemented on the data from a survey in Metropolitan Toronto of attitudes toward smoking in the workplace. Using a Bayesian model, the posterior distribution of the model parameters is sampled by Markov Chain Monte Carlo methods. The results reveal that the non-response is not ignorable and those who do not respond are twice as likely to favor unrestricted smoking in the workplace as are those who do.

KEY WORDS: Call-backs, number of; Bayesian analysis; Markov Chain Monte Carlo method; Informative non-response; Ignorable non-response.

1. INTRODUCTION

Given the reality of non-response in every survey, it is of interest to determine how to account for this non-response in the interpretation of the collected data. Rubin (1976) gives necessary and sufficient conditions for such an analysis to be identical from, respectively, a frequentist, likelihood, and Bayesian perspectives, to an analysis based on a model incorporating a missingness mechanism. Building on this, Little and Rubin (1987) led to an extensive literature modeling non-response in an informative, non-ignorable way.

Information about the interaction between the survey and the surveyed can sharpen the analysis of the import of missing data in a survey. The example in this paper concerns the attitudes of Toronto citizens about smoking in the workplace. Random telephone numbers were chosen; at least twelve calls were made to try to reach the intended respondents. Our data for the respondents includes only the number of calls until the survey was completed, not the timing of the unsuccessful calls. With even this attenuated data on how difficult the respondent was to reach, we find our view of the results of the survey to be importantly informed by the number of unsuccessful calls.

The use of information on the number of calls to a subject chosen to participate in a survey is not unique. Potthoff, Manton and Woodbury (1993) present a method for correcting for survey bias due to non-availability by weighting based on the number of call-backs. While our analysis also focuses on the bias due to non-availability, there are major differences. Instead of assuming that refusals do not exist, we allow for and utilize their potential existence in modeling the mechanism which causes non-

response. In the analysis that follows, the relationship of non-response to the response variable of interest in the survey is evaluated along with other explanatory variables, after weighting for both household size and the appropriate population demographics. In doing so we address not only whether error exists due to non-availability, but also whether stratification of the respondents by household size and the then current age/sex distribution may eliminate the necessity for accounting for the error by the introduction of a mechanism which describes the non-response. Note that here we match the groupings of Pederson, Bull and Ashley (1996) used in the original published analyses of the dataset; more complex cell adjustment procedures are possible (*e.g.*, Little 1996; Eltinge and Yansaneh 1997, and references cited therein).

The remainder of this article is organized as follows: Section 2 gives more detail on the survey; section 3 introduces the methodology employed; Sections 4 and 5 respectively explore missing-at-random and non-ignorably-missing models; Section 6 discusses the priors distributions chosen for the main analysis, whose results are explained in section 7. Finally, section 8 gives our conclusions.

2. THE SURVEY

A bylaw regulating smoking in the workplace in the City of Toronto took effect on March 1, 1988. From January 1988 to the present, a series of six surveys have been conducted to assess attitudes of the public toward smoking, awareness of health risks related to smoking, and the impact of the law on the residents of Metropolitan Toronto. The data being utilized in this analysis comprises the third phase

¹ Louis T. Mariano is a Ph.D. candidate, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; Joseph B. Kadane is Leonard J. Savage University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

of this series. Northrup (1993) provides the technical documentation for this survey. For clarity, when necessary, the data being analyzed here is referred to as the Phase III data, and information from the first two surveys is referred to as the Phase I & II data.

Northrup (1993) indicates that the data of interest, which were made available by the Institute for Social Research (ISR) at York University, were collected from 1,429 residents of the Metropolitan Toronto area in December 1992 and March 1993. A two-stage probability selection process was utilized to select survey respondents. The first stage employed random digit dialing. The second stage used the most recent birthday method to select one adult individual once an eligible residence was reached. The responses were then weighted by the number of adults in the household. In the analysis that follows, post-stratification weighting was also applied to the census age-sex distribution to adjust for the underrepresentation of some population subgroups. The number of distinct phone lines in the household was not taken into consideration during the data collection.

The number of calls it took to reach each respondent is included as a variable in the dataset, and there are no missing values for this variable. Northrup (1993) explains that the 1,429 responses came from a sample of 5,702 telephone numbers generated by the random digit dialing method. Of these numbers, 2,286 were verified to be eligible households, and 3,150 of the numbers in the sample were not eligible. The status of the remaining 266 numbers was not able to be determined. It has been assumed by ISR that the household eligibility rate of these 266 numbers was equal to the rate for the rest of the sample. This eligibility rate implies an estimated total of 2,398 households in the sample and a response rate of 60%. Thus, an estimated 969 subjects chosen to participate in the survey did not respond. Each subject received a minimum of 12 calls, including day, night, and weekend calls, before being classified as non-respondent.

The dependent variable, for the purpose of this analysis, is an individual's opinion on the regulation of smoking in the workplace, in one of three categories. Category "0" indicates smoking should be permitted in restricted areas only, category "1" indicates smoking should not be permitted at all, and category "2" indicates smoking should not be restricted at all. For each subject chosen to participate in the survey, let $Y_i \in \{0, 1, 2\}$ represent the opinion of subject i .

The data comprises of the answers to 50 survey questions as well as 18 other variables identifying characteristics of the subject. Included in these are:

- "K-risk" is an integer score from 0 to 12 which indicates knowledge of the risks and effects of second-hand smoke.

- "Smoker" indicates the smoking status of the subject: "Current smoker" (S), "Former smoker" (SQ) or, "Never smoked" (NS).
- "Bother" indicates if second-hand smoke bothers the subject: "Always bothers" (b.A), "Usually bothers" (b.USUL), or "Does not bother" (b.NO).
- "Age": $(\text{Age in years} - 50) / 10$.

Pederson, Bull, Ashley and Lefcoe (1989) created a "Knowledge of health effects score" on passive smoking out of the answers to six survey questions, which measured a subject's knowledge of the effects of second-hand smoke. Pederson *et al.*'s questions were used in Phase III to create their score, here renamed "K-risk". A higher K-risk score indicates a greater knowledge of the risks of second-hand smoke. The variable "Age" was shifted and rescaled to match how age was treated by Bull (1994) in the Phase I & II analysis.

3. OVERVIEW OF METHODOLOGY

The fundamental question of interest is: "May we ignore the unit non-response and treat the observed data as a random subsample of the population?" Mapping to the terminology of Little and Rubin (1987) and Rubin (1976): If we may treat the observed data for the dependent variable of interest as a random subsample, we call the missing data "missing completely at random" (MCAR). If we may treat the observed data for the dependent variable of interest as a random subsample, after conditioning on the explanatory variables, we call the missing data "missing at random" (MAR). Let θ represent the parameters of the data and let π represent the parameters describing the missing data process. Rubin (1976) calls the parameters π and θ distinct "if there are no *a priori* ties, via parameter space restrictions or prior distributions, between π and θ ." If either the MCAR or MAR cases apply and if π and θ are distinct, the mechanism which causes the missing data is said to be "ignorable" for inference about the distribution of the variable of interest. If the missing data for the dependent variable of interest is dependent on the values of that data, then the mechanism which causes the missing data is said to be "non-ignorable" (NI). Groves and Couper (1998) note that when the likelihood of participation is a function of the desired response variable, the non-response bias can be relatively high, even with a good response rate.

Let R_i be an indicator of response. $R_i = I_{\{\text{respondent}\}}(\text{subject } i)$ and $R = (R_1, \dots, R_n)^T$. Little and Rubin (1987) suggest that one possible method for accounting for the non-response mechanism is to include this response indicator variable in the model. We may call the mechanism which causes the missing data ignorable if π and θ are distinct and:

$$f(R | Y_{\text{obs}}, Y_{\text{mis}}, \pi) = f(R | Y_{\text{obs}}, \pi) \quad (1)$$

where Y_{obs} and Y_{mis} represent the observed and missing portions of the dependent variable of interest.

The terms "MAR assumption" and "NI assumption" will be used throughout this analysis. For clarity, the term "MAR assumption" is defined as the assumption that the missing data mechanism is ignorable for inference with respect to the dependent variable identified in section 2. That is, the observed values of that variable are a random subsample of the population, possibly within poststrata, and it is not necessary to account for the missing data mechanism. The term "NI assumption" is defined as the assumption that the missing data mechanism is non-ignorable and the data collected for the dependent variable of interest cannot be treated as a random subsample. Specifically, inference for the population must involve the missing data mechanism.

The approach to assessing the MAR assumption is comprised of three steps. The first step is the examination of what one might do under the MAR assumption. Since the dependent variable of interest has three categories and some of the explanatory variables are quantitative, polytomous logistic regression is employed. Both frequentist and Bayesian forms of the logistic regression model are examined.

In the second step, and NI model is constructed. The non-response mechanism is modeled utilizing the information available about the number of calls made to each subject. Here, the idea of a surviving fraction in the sample is examined to model whether it is actually possible to reach all the intended respondents. Then, the non-response mechanism is related to the dependent variable by including the number of calls in the logistic regression model.

In the development of the NI model, we employ a Bayesian approach to allow for an examination of the values the missing data are likely to take, given the observed data and the model parameters. This is accomplished by utilizing a data augmentation approach, where the missing data are imputed in each iteration of a Markov Chain Monte Carlo (MCMC) simulation. A possible alternative would be to utilize the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin 1977) to compute the maximum likelihood estimates (MLE's) of the missing values.

In the third step, an evaluation of the MAR assumption is made. Non-zero coefficients for the number of calls in the logistic regression portion of the NI model will imply that the number of calls does make a difference; *i.e.*, the opinions of those who did not respond in the first 12 calls are likely to differ from those who responded in just a small number of calls. In this case, the missing data mechanism is not independent of the values of the missing data and an MAR assumption would be inappropriate. Next, the log odds of response among the three models are examined. Differences here identify the magnitude of the error that a faulty MAR assumption causes. So, in the evaluation of the MAR assumption, the questions "is there a difference?" and "how large is the difference?" are both addressed.

4. MAR MODELS

4.1 Logistic Regression

Using the data collected from the ($m = 1,429$) subjects that did respond to the survey, weighted logistic regression was employed to model the public's opinion on smoking in the workplace. The collection of candidate predictors found in the survey questions and the background information was narrowed utilizing a series of Wald tests. Then likelihood ratio tests, AIC, and BIC were used to compare the possible models. The model with the best fit was found to be the one which included additive terms for the variables "K-risk", "Smoker", "Bother", and "Age", as defined in section 2.

As each of the models examined in this analysis employs a logistic regression component, it is useful here to illustrate the notation being used. Category "0", "smoking allowed in restricted areas only" was chosen to be the reference category. Recall $Y_i \in \{0, 1, 2\}$. For the MAR model, we use only the observed values of the subject's opinion on workplace smoking, $Y_{\text{obs}} = (Y_1, \dots, Y_m)$. Let $Y_{ij} = I_{\{j\}}(Y_i)$ be an indicator of subject i responding in category j , and let W_i represent the weight each subject received. As in the original published analyses of this dataset (Pederson *et al.* 1996) both household (see Northrup 1993) and post-stratification (see Appendix A) weighting were used in the consideration of all models here.

The two categorical explanatory variables, "Smoker" and "Bother", were included in the model by utilizing indicator variables for two of the three categories, with the effect of the third category being absorbed in the intercept term. For "Smoker", " S_i " and " SQ_i " were included as indicators that subject i was either a current smoker or a smoker who had quit. For "Bother", " $b.USUL_i$ " and " $b.NO_i$ " were included as indicators that second had smoke usually bothered or did not bother subject i .

Let X_i = represent the vector for explanatory variables for subject i . Then,

$$X_i = (K\text{-risk}_i, S_i, SQ_i, b.USUL_i, b.NO_i, \text{Age}_i).$$

Here we use an unordered multinomial logit model to consider $p_j(x_i) = P(Y_{ij} = 1 | X_i = x_i)$, the probability that subject i responds in category $j \in \{0, 1, 2\}$, given the observed explanatory variables for subject i . This model, of course, utilizes linear equations η_{ij} describing the log odds of subject i responding in category j versus the reference category $j = 0$. So, for $j = 1, 2$ we wish to examine:

$$\ln \frac{p_j(x_i)}{p_0(x_i)} = \eta_{ij} = \beta_{0j} + X_i \beta_j, \quad (2)$$

with $\eta_{i0} = 0$. The two resultant linear equations, η_{i1} and η_{i2} , each have seven coefficients, including an intercept term β_{0j} and those displayed below:

$$\beta_j = (\beta_{K\text{-risk}_j}, \beta_{S_j}, \beta_{SQ_j}, \beta_{b.USUL_j}, \beta_{b.NO_j}, \beta_{\text{Age}_j}).$$

The MAR logistic regression model has 14 parameters. The vector of these 14 parameters, represented by $\beta = (\beta_{01}, \beta_{11}, \beta_{02}, \beta_{21})$ has the likelihood (or, more appropriately, pseudo-likelihood, since the weights are incorporated through the variable W_i):

$$L(\beta) \propto \prod_{i=1}^m \prod_{j=0}^2 \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \quad (3)$$

4.2 Bayesian Logistic Regression

The likelihood in equation (3) and the data collected from the survey respondents are utilized in the Bayesian analysis. The same four explanatory variables selected in the frequentist analysis above are used as the explanatory variables here. Prior distributions, discussed in section 6, were assigned to the logistic regression parameters. An MCMC simulation is utilized in order to draw from the posterior distribution of the parameters.

5. NI MODEL

5.1 Modeling the Non-Response Mechanism

Since the missing values are not necessarily missing at random, the mechanism which caused them to be missing must be addressed. Northrup (1993) indicates that non-respondent subjects chosen to participate in the survey were called a minimum of 12 times, including a minimum of three day, four evening and four weekend calls. Unfortunately, other useful information regarding the number of calls was not retained. We do not know which of the non-respondents were called more than twelve times or whether an individual call was placed during the day, evening, or weekend. We also are unaware of the details of the non-response, such as whether the subject was contacted but

refused to participate, whether the calls were ever answered by a machine, or whether they were answered at all. Thus, stratification of the non-respondents was not possible, and they were all treated as exchangeable in this analysis.

Each subject was called a number of times until the survey was successfully completed or they were classified as non-respondent. For the respondents, the number of calls variable (C_i) describes the number of trials until the first success for subject i . Thus, one might expect the number of calls to follow a Geometric distribution with truncated observations for the non-respondents. Specifically, let $\pi = P(\text{a call is successful})$; then, consider $C_i \sim \text{Geometric}(\pi)$ and $P(C_i = c_i) = \pi(1 - \pi)^{c_i - 1}$. Note that if auxiliary information about the number of calls to the non-respondents were available (e.g., Groves and Couper 1998), we could have also considered conditional response probabilities here.

The histograms in Figure 1 compare the data (through the first twelve calls) to a Geometric distribution with parameter $\pi = .225$, which appears to match fairly well. The sample order statistics suggest $\pi \in (.2, .25)$. The histogram of the actual survey data reveals that the number of subjects reached on the first call are fewer than the number reached on the second call. It is possible that more of the second calls were placed at a time which had a higher success rate.

Suppose $\pi = .225$; by the memoryless property of the Geometric distribution, we would expect 218 of the 969 non-respondents to reply on the 13th call. This would make the data through the first 13 calls appear as in Figure 2. Clearly, Figure 2 does not display the behavior of a Geometric random variable. Consider the following question: "If all subjects were called an unlimited amount of times, would they all have been reached?" Answering "yes" to that question for this dataset results in the problem illustrated in Figure 2.

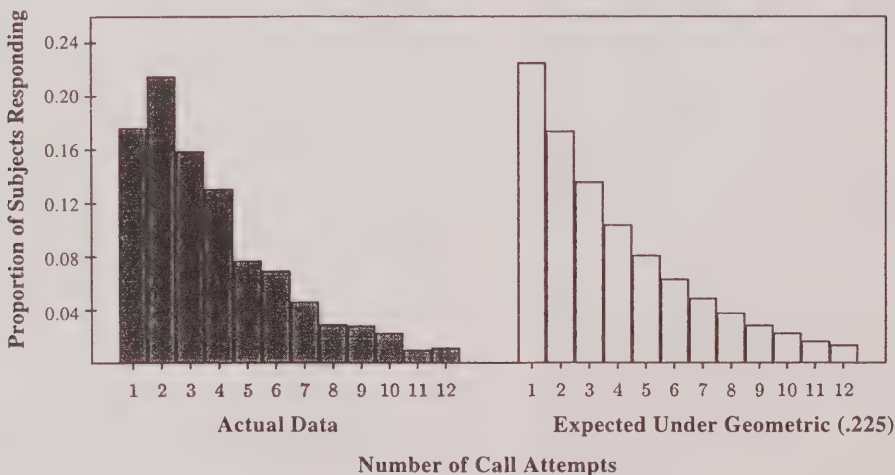


Figure 1. Comparison of the actual survey data for successful calls in the first 12 attempts to expected results based on a Geometric (.225) distribution for the number of calls needed to complete the survey.

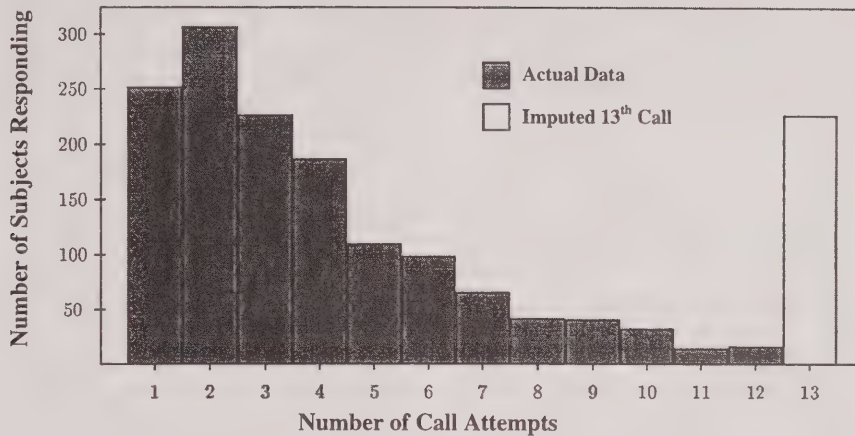


Figure 2. Display of the actual number of successful calls on each attempts through the first 12 and the expected number of successful calls on the 13th attempt. The expectation for the 13th call is based on a Geometric (.225) distribution to model the number of calls until the survey is completed

Given the information outlined above, the assertion that “not all subjects chosen for the survey are reachable” is a viable one. Maller and Zhou (1996) discuss immune subjects – individuals who are not subject to the event of interest. Following their terminology, if it is not possible to procure a response from a subject chosen for the survey given an unlimited amount of calls, that subject is categorized as immune. Subjects who are not immune are categorized as “susceptible”. The set of immune (*i.e.*, non-susceptible) subjects comprise the “surviving fraction” of the sample. Mapping to more familiar terminology, the immune subjects include those who were reached and refused, those who would have refused if they had been reached, and those cases of a physical or mental inability to ever participate. Northrup (1993) indicates that those who initially refused to participate were subsequently contacted by the most senior interviewers, so, we make the assumption here that all remaining refusals would not ever participate. The susceptible group includes the respondents, those who would have responded if successfully contacted, and those who were physically or mentally unable to participate during the data collection period but were willing and able at some other time.

Let the variable $Z_i = I_{\{\text{subject } i \text{ is susceptible}\}}$ (subject i) be an indicator of the susceptibility of subject i , and $\rho = P(\text{subject } i \text{ is susceptible})$, *i.e.*, $Z_i \sim \text{Bernoulli}(\rho)$. Now suppose that the number of calls to the susceptible subjects follows a Geometric distribution, *i.e.*, $C_i | Z_i = 1 \sim \text{Geometric}(\pi)$. Does this eliminate the problem illustrated in Figure 2?

Let R_i be an indicator of response of subject i . The non-response mechanism can be accounted for by including these response indicators in the model. However, the introduction of the susceptibility variable implies two distinct

classes of non-response. So, it is possible to be more detailed and use both the susceptibility $Z = (Z_1, \dots, Z_n)^T$ and the response R indicators in a mixture model describing the non-response. Updating Equation (1), the missing data mechanism is ignorable if and only if (π, ρ) is distinct from θ and

$$f(R, Z | Y_{\text{obs}}, Y_{\text{mis}}, \pi, \rho) = f(R, Z | Y_{\text{obs}}, \pi, \rho). \quad (4)$$

Let $C_{\text{obs}} = (C_1, \dots, C_m)$ and $Z_{\text{obs}} = (Z_1, \dots, Z_m)$ be the vectors of the number of calls and the observed susceptibility for each respondent. Also, let $R = (R_1, \dots, R_n)$ be the vector of response for each intended respondent. Every subject, i , may be classified by response into three mutually exclusive groups, A_{obs} – observed, A_{mis} – missing, and A_{imm} – immune, where:

$$A_{\text{obs}} = \{i: i \text{ was Susceptible and Responded}\}$$

$$A_{\text{mis}} = \{i: i \text{ was Susceptible but did not Respond in 12 calls}\}$$

$$A_{\text{imm}} = \{i: i \text{ was not Susceptible}\}.$$

The probability that a subject is in each of these categories may be calculated as follows:

$$P(i \in A_{\text{obs}}) = P(Z_i = 1, R_i = 1, C_i = c_i) = \rho \pi (1 - \pi)^{c_i - 1}$$

$$P(i \in A_{\text{mis}}) = P(Z_i = 1, R_i = 0, C_i > 12) = \rho (1 - \pi)^{12}$$

$$P(i \in A_{\text{imm}}) = P(Z_i = 0) = 1 - \rho.$$

The data indicates $m = 1,429$ subjects in A_{obs} and $n - m = 969$ non-responsive subjects in $A_{\text{mis}} \cup A_{\text{imm}}$; $n = 2,398$ is the estimated total number of subjects chosen to participate in the survey. Thus, the joint density of Z_{obs}, R and C_{obs} given ρ and π is:

$$f(Z_{\text{obs}}, R, C_{\text{obs}} | \rho, \pi) \propto$$

$$\left[\rho^m \pi^m (1 - \pi)^{(\sum_{i=1}^m c_i) - m} \right] \times \left[(1 - \rho) + \rho (1 - \pi)^{12} \right]^{n - m} \quad (5)$$

The mixture model described by Equation 5 may be viewed as a special case of the non-response models discussed in Drew and Fuller (1981).

It would be useful to confirm that the above joint distribution accurately represents the response pattern of the susceptibles in the dataset. The MLE estimate for ρ is simply the proportion of respondents in the sample, which clearly underestimates ρ . Setting $U(0, 1)$ prior distributions for both ρ and π and examining their joint posterior distribution by MCMC simulation, the posterior medians are found to be $\rho = .636$ and $\pi = .205$, with equal-tailed posterior credible intervals of (.613, .659) and (.191, .219) for ρ and π respectively. Figure 3 illustrates how the dataset might look after imputing the missing number of calls for our susceptible non-respondents based on these posterior medians. The problem previously displayed in Figure 2 has now been mostly eliminated.

While the Geometric distribution appears sufficient (after accounting for susceptibility), a referee questions the use of the Geometric distribution as it does not make use of possibly useful covariates. As explained above, the covariates we think would be most useful for this purpose were

not collected. One alternative for modeling the response mechanism of the susceptibles is to use a discretized Gamma distribution. In cases where more complexity is necessary, the v-Poisson (a two parameter Poisson which generalizes some well known discrete distributions, including the Geometric) of Shmueli, Minka, Kadane, Borle and Boatwright (2001) may also be considered.

5.2 Relating Non-Response to the Dependent Variable – The NI Model

Since the non-response of the susceptibles is described by the conditional Geometric distribution of the number of calls, the effect of the non-response of the susceptibles on the dependent variable may be considered by including the number of calls as an additional explanatory variable in the logistic regression likelihood. This will create two additional parameters in the logistic regression portion of the model, which are the coefficients of the number of calls, β_{call_i} in each of the linear equations η_{ij} described in equation (2).

Non-zero coefficients for the number of calls, then, would indicate that the dependent variable is not independent of the non-response mechanism, and, hence the non-response mechanism is non-ignorable. If these coefficients are zero, the non-response of the susceptibles is ignorable. Conclusions made here rely upon the underlying modeling assumption that the relationship among the number of calls, the dependent variable and the other explanatory variables considered is the same for the respondents and susceptible non-respondents. Including the number of calls in the logistic regression portion of the model does not address the immune subjects, since there will never be the realization of a successful call to them.

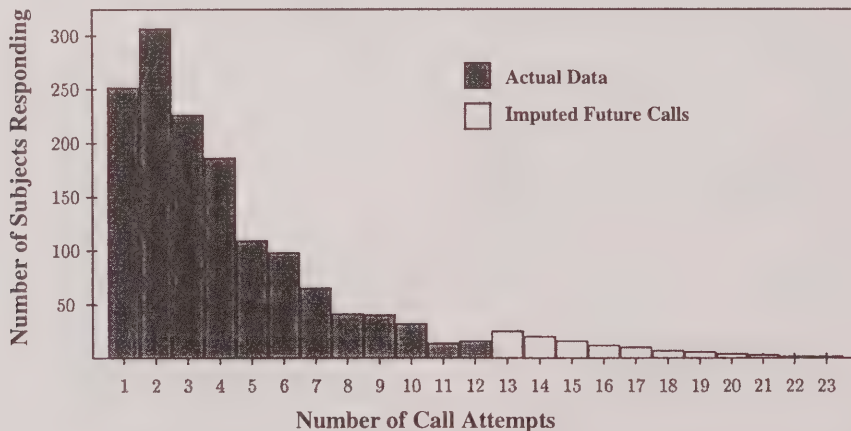


Figure 3. Display of the actual number of successful calls on each attempt through the first 12 and the expected number of successful calls for call attempts 13 and higher. Imputed values are based on a probability of a successful call of .205 and a probability of susceptibility of .636.

The full pseudo-likelihood for the NI model (or, more precisely, the susceptible NI model) is the product of the non-response and logistic regression pieces:

$$L(\rho, \pi, \beta) \propto \left[\rho^m \pi^m (1 - \pi)^{(\sum_{i=1}^m c_i) - m} \right] \times \left[(1 - \rho) + \rho (1 - \pi)^{12} \right]^{n-m} \times \left[\prod_{i=1}^m \prod_{j=0}^2 \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \right]. \quad (6)$$

Note that the household and post-stratification weighting variable W_i is included here in an effort to account for whether proper stratification of the respondents may eliminate the necessity for the introduction of a mechanism to describe non-response.

5.3 Data Augmentation

Tanner and Wong (1987) suggest an iterative method for computation of posterior distributions when faced with missing data. This method applies whenever augmenting the dataset makes it easier to analyze and the augmented items are easily generated. Consider the following additional notation: Let S represent the total number of susceptible subjects in the sample. $S = \sum_{i=1}^n Z_i$, $S \sim \text{Binomial}(\rho)$. Let X be the matrix of explanatory variables (including the number of calls) for all the subjects selected to participate in the survey. Let $Y = (Y_1, \dots, Y_n)$ be the vector of their responses. Partitions X into $\{X_{\text{obs}}, X_{\text{mis}}, X_{\text{imm}}\}$ and Y into $\{Y_{\text{obs}}, Y_{\text{mis}}, Y_{\text{imm}}\}$. Also, by the memoryless property of the Geometric distribution, the distribution of the additional number of calls required to reach the subjects in A_{mis} is known, and may be expressed: $\forall i \in A_{\text{mis}}$, let $V_i = C_i - 12$, which is also distributed as a Geometric random variable with parameter π .

Now suppose that the true values of S , X_{mis} , and Y_{mis} were known. The likelihood could then be considered in the form:

$$L(\rho, \pi, \beta \mid X_{\text{obs}}, X_{\text{mis}}, Y_{\text{obs}}, Y_{\text{mis}}, S, R) \propto \left[(\rho \pi)^s (1 - \pi)^{(\sum C_{\text{sus}}) - s} \right] \times \left[(1 - \rho)^{n-s} \right] \times \left[\prod_{i=1}^s \prod_{j=0}^2 \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \right], \quad (7)$$

where $\sum C_{\text{sus}} = \sum C_{\text{obs}} + \sum (V_i + 12)$ is the number of calls that would have been necessary to reach all susceptibles and the summands are taken over the appropriate range of subjects.

Although the true values of S , X_{mis} , and Y_{mis} are unknown, one may utilize what is known about the behavior of these variables to impute stochastically possible values for them within the MCMC algorithm. Given ρ , a value for S may be drawn from a truncated Binomial (2,398, ρ), where $1,429 \leq S \leq 2,398$. Given S , the number of subjects in A_{mis} is known. For each of these subjects in A_{mis} a value $V_i \sim \text{Geometric}(\pi)$ may be drawn, which results in an imputation for the number of calls needed to reach each susceptible but unreached subject. The relationships among the number of calls and the other explanatory variables may then be exploited to impute values for the rest of X_{mis} . Specifically, the missing values of Age and K-risk are imputed by regressing Calls on Age and K-risk respectively and predicting from the resultant linear equations. Similarly, the missing values of Smoker and Bother are imputed via logistic regression on each, using Calls as the explanatory variable. Here the model assumptions are checked using the respondents data, and an assumption is being made that these same relationships hold for the susceptible non-respondents. Note that these regression and logistic regression equations are fit in the Bayesian context (e.g., Gelman, Carlin, Stern and Rubin 1998) and necessitate the inclusion of additional parameters, β_j , in the MCMC process which describe these relationships (see Appendix B for more detail). We chose this imputation plan in the interest of the efficiency of the full MCMC algorithm. An alternative would be to impute the missing values for a particular explanatory variable conditional on all the remaining variables (e.g., Rubin 1996). Finally, Y_{mis} may be predicted by utilizing the imputed values of X_{mis} and the relationship described in the logistic regression model. In the interest of the exchangeability of the susceptible non-respondents in the absence of subsequent stratification information, we apply a weight of 1.0 to all the imputed Y_{mis} values; an alternative here would be to impute the sex and household size of the susceptible non-respondents, in addition to their age, and apply the weighting procedure described in Appendix A to the imputed Y_{mis} .

5.4 Sampling from the Posterior Distribution

The full MCMC simulation consists of a Metropolis algorithm supplemented in every iteration with the data augmentation described above. An outline of the MCMC algorithm used may be found in Appendix B. Convergence was assessed utilizing the method of Hiedelberger and Welch (1983) as described in Cowles and Carlin (1996). MacEachern and Berliner (1994) assert that, under loose conditions, subsampling the MCMC simulated values to account for autocorrelation will result in poorer estimators. Following their suggestion, all simulated values, after an appropriate burn-in period, were used in the analysis that follows.

6. CHOICE OF PRIOR DISTRIBUTIONS

In the evaluation of possible prior distributions for the parameters of both the NI and MAR models, the goal of the comparison of the various models was taken into consideration. The choice of prior distributions for the parameters was made from the perspective of the MAR belief. Two possibilities were examined.

The first option is built around the utilization of the Phase I & II surveys. Since these surveys were similar to and were completed prior to the Phase III survey which comprises our data, information contained in these first two surveys may be utilized in the construction of priors. The same dependent variable was contained in the Phase I & II dataset, along with the variables Smoker, Age, and K-risk. A logistic regression model was compiled from the Phase I & II data to describe the relationship between the opinion on workplace smoking and these three explanatory variables. Normal priors were constructed for the coefficients of these three variables centered at their MLE's, but with increased standard error. The error terms were increased due to three factors:

- i) There was a three year span between the Phase II and Phase III surveys; opinions may have changed over that time, possibly as a result of the impact of the bylaw.
- ii) The MLE's were calculated under the same MAR assumption being evaluated.
- iii) Prior to the collection of the Phase III data, there existed the possibility that other explanatory variables would be included in the model; in the presence of other variables, the effect of these three could be altered.

Although the variances were increased, the means were not changed, since it was unknown, *a priori*, in what direction any change might occur. Since the available Phase I & II data contained no information about the Calls or Bother variables, the coefficients of these were assigned a diffuse Normal (0,9) prior. For clarity, this option will be referred to as the "Phase I & II prior" in this analysis.

In the second option Normal (0,9) priors are assigned to each of the logistic regression coefficients. One motivation for this choice is that, for the same three reasons the error terms were increased above, the variables common to the Phase I & II and Phase III surveys are not exchangeable. Thus, construction based on the Phase I & II results would be inappropriate. This option will be referred to as the "Central prior".

The choice to use Normal (0,9) distributions here is for convenience. Centering the prior at zero gives equal weight to either direction of the relationship. We believe the choice of a variance of nine to be adequate without being overly diffuse. The use of improper priors could lead to a Markov Chain Monte Carlo simulation that never converges, and, as Natarajan and Kass (2000) show, an overly diffuse proper prior may behave like an improper one. In section (7.2), we

offer a sensitivity analysis to evaluate how the results are effected by the choice of prior.

The non-response parameters of the NI model, ρ and π , were treated the same under both prior options. There was no additional information available about the probability of a successful call or the probability of susceptibility. Thus, ρ and π were each assigned a $U(0,1)$ prior.

The data augmentation parameters found in each of the logistic regression equations, β_p , were independently given diffuse Normal (0,9) priors. For each linear regression equation found in the data augmentation process, the coefficients, β_p , and variance, σ_p^2 , were set to $p(\beta_p, \sigma_p^2) \propto 1/\sigma_p^2$, the standard non-informative prior distribution (e.g., Gelman *et al.* 1998). Note that the closed forms of the posterior distributions of the linear regression parameters are known and may be drawn from directly.

7. RESULTS

First, the validity of the MAR assumption is examined through the coefficients of the number of calls variable. Then, the NI model is evaluated with respect to sensitivity to the choice of prior. Finally, the magnitude of the impact of a faulty MAR assumption for this dataset is investigated by illustrating the change in the odds of response.

7.1 Coefficients for the Number of Calls

For both the Phase I & II and Central priors, Figure 4 displays the posterior density (solid line) and 95% credible interval estimates (dotted lines) of the coefficient of the calls variable in η_{11} in the NI model, and compares them to the point $\beta_{\text{call}_1} = 0$ (dashed lines). The results clearly indicate this coefficient differs from zero. We also find a non-zero result in η_{12} , where, using the Phase I & II prior, the 95% HPD credible interval for β_{call_2} is (-0.03613, 0.11595).

The non-zero coefficient of C_i demonstrates a dependence between the number of calls and the subject's opinion on smoking in the workplace. Thus, the dependent variable and the non-response mechanism are not independent under the conditions discussed in section 5.2. This results implies that an assumption that the missing observations are missing at random prior to accounting for the non-response mechanism is incorrect for this dataset.

There is a hint in Figure 3 that the probability of a successful call decreases as the call number increases. To verify the assumption that the relationship between the number of calls and the log odds of response is linear, a second Bayesian NI model was constructed. This model split the calls variable into two, $C_i I_{(C_i < 7)}$ and $C_i I_{(C_i \geq 7)}$, based on whether the number of calls were fewer than seven. The posterior distributions of the coefficients of these two variables were then compared and evidence that they are essentially different was not found. In particular, for η_{11} the 95% credible interval for $C_i I_{(C_i \geq 7)}$ contained the same interval for $C_i I_{(C_i < 7)}$, and for η_{12} the 95% credible intervals strongly overlapped.

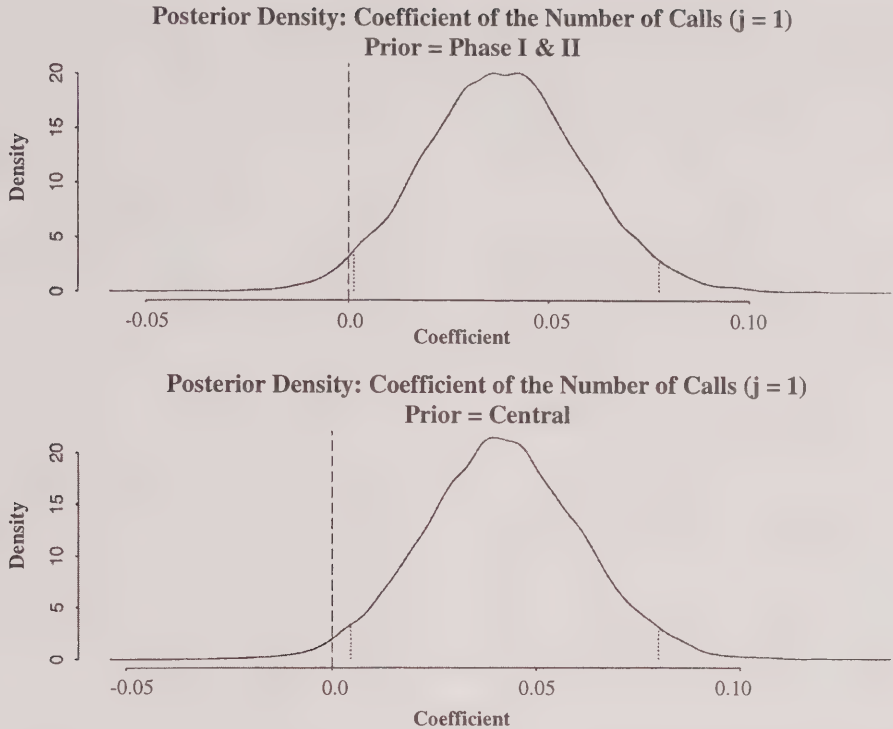


Figure 4. Display of β_{call_i} , the coefficient of the calls variable in η_{i1} : posterior density (solid line) and 95% equal tailed credible interval (dotted line), compared to $\beta_{\text{call}_i} = 0$ (dashed line).

7.2 Sensitivity to Priors

Would different prior distributions, either on the calls coefficient or on the others, make a difference in the effect illustrated above? Table 1 displays 95% HPD credible intervals for the coefficient of the calls variable in the first logit equation of the NI model for six different priors. The priors include the Phase I & II and Central priors as well as four others - labeled options 3, 4, 5, and 6. Options 3 and 4 resemble the Central prior except that they change the prior distribution on the coefficient of the number of calls to Normal (1,9) and Normal (-1,9) respectively. Option 5 places Normal (0,9) priors on β_{call_i} , β_{age_i} , and β_{b,USUL_i} , a Normal (1,9) prior on β_{01} , a Normal (.5,9) prior on $\beta_{K-\text{risk}_i}$, a Normal (-1,9) prior on β_{S_1} and Normal (-5.9) priors on β_{SQ_1} and β_{b,NO_1} . Option 6 takes the Central Prior and reduces all the variances from nine to two.

Under all six priors, Table 1 demonstrates that the coefficient of the calls variable in the first logit equation clearly differs from zero. The finding that the missing data mechanism is non-ignorable for this dataset does not appear to be effected by the choice of prior among these options.

Table 1
95% HPD Credible Intervals for β_{call_i} Under six Different Prior Distributions

Prior	Coefficient of the number of Calls " C_i " in η_{i1}	
	95% intervals	
	Lower Bound	Upper Bound
Phase I & II	0.00129	0.07746
Central	0.00446	0.07980
Option 3	0.00447	0.07983
Option 4	0.00441	0.07975
Option 5	0.00440	0.07970
Option 6	0.00436	0.07944

7.3 Effect on Odds of Response

Given the failure of the MAR assumption shown above, it is of interest to question the relevance of the error that using the MAR assumption would create. The magnitude of the error induced by a faulty MAR assumption may be illustrated by examination of its effect on the odds ratio $p_1(x_i)/p_0(x_i)$. First, we consider the effect on a typical

respondent profile. The modal respondent was a non-smoker between the ages of 25-35 years old who was usually bothered by second-hand smoke, had a K-risk of 11 and could be reached in 2 calls. We label this modal respondent as Subject 1. Table 2 demonstrates the change in posterior odds for Subject 1 when called 13 times.

Table 2

Comparison of the Odds of Response for 4 Typical Subjects. Posterior Medians Were Used As the Point Estimates for the Coefficients in the Bayesian Models; the MLE Was Used for the Frequentist Model

	Subject 1	Subject 2	Subject 3	Subject 4
Smoker	No	No	Former	Yes
Age	30	50	27	40
Bother	Usually	Always	No	No
K-risk	11	12	7	3
Model	Odds $Y=1/Y=0$			
MAR MLE	0.674	2.105	0.457	0.396
MAR Phase I & II prior	0.703	4.487	0.209	0.116
NI Phase I & II prior: 2 calls	0.640	4.024	0.202	0.108
NI Central prior: 2 calls	0.593	4.442	0.162	0.102
Option 3: 2 calls	0.594	4.449	0.162	0.102
Option 4: 2 calls	0.592	4.435	0.162	0.101
Option 5: 2 calls	0.590	4.423	0.161	0.101
Option 6: 2 calls	0.590	4.426	0.161	0.101
NI Phase I & II prior: 13 calls	0.974	6.128	0.308	0.165
NI Central prior: 13 calls	0.936	7.013	0.256	0.160
Option 3: 13 calls	0.937	7.026	0.256	0.161
Option 4: 13 calls	0.934	7.000	0.255	0.160
Option 5: 13 calls	0.930	6.975	0.254	0.159
Option 6: 13 calls	0.931	6.980	0.254	0.160

The Subject 1 column Table 2 indicates a dramatic difference in the posterior odds when the non-response mechanism is taken into consideration. For this typical respondent profile, when the number of calls is increased from two to thirteen the posterior odds of choosing "Smoking should not be permitted at all" over "Smoking should be permitted in restricted areas only" increases by 52.18% under the Phase I & II prior and 57.84% when using the Central prior. This is dramatic evidence of the relationship between the dependent variable and the non-response mechanism.

Are the results for the modal subject above typical? Table 2 also displays the effects on the odds of response under the NI model for three additional test subject profiles for each of the six different priors considered above. Subject 2 is a fifty year old non-smoker who is always bothered by smoke and has a perfect "K-risk" score. Subject 3 is a 27 year old former smoker who is not bothered by smoke and has a "K-risk" score of seven. Subject 4 is a 40 year old smoker who is not bothered by smoke and has a "K-risk" score of three. On multiple subjects with multiple priors, Table 2 consistently shows

the same result. Increasing the number of calls to greater than 12 will increase the posterior odds of choosing category "1" over category "0". For each of the test subjects and priors found in Table 2, the increase was between 52.18% and 58.41%.

Similar results were found when examining the odds of choosing the "Smoking should not be restricted at all" category over the "Smoking should be permitted in restricted areas only" category. Using test subjects which were a current and a former smoker (Subjects 3 and 4 above), the posterior odds increased 46.7% when the number of calls was increased from 2 to 13 under the Phase I & II prior.

7.4 Effect on Probability of Response

With the shift in posterior odds illustrated above comes a corresponding shift in the estimated probabilities that a subject will respond in a particular category. Among the respondents, 57.45% chose category "0", 40.64% chose category "1", and 1.91% chose category "2". The number of non-respondent susceptibles have a posterior median of 469, with a 95% credible interval of (25, 944). On average, 55.88% of the simulated non-respondent susceptibles chose category "0", 40.03% chose category "1", and 4.08% chose category "2". While, for categories "0" and "1", the average values for the non-respondent susceptibles do fall within the 95% confidence intervals for the proportions of the respondents in these categories, the point estimates for each category shift when the non-response mechanism is included in the model. In comparing the category "2" results, we estimate that non-respondents are twice as likely to favor no restrictions on smoking (category "2") than are respondents. While the low number of subjects found in category "2" are unlikely to provoke a change in workplace smoking law, the increasingly noted in the non-respondents in this category serves as an example of how the lack of proper consideration of the non-respondents could lead to flawed conclusions about the data.

8. CONCLUSION

Section 7 demonstrates that, for the dependent variable of interest in this dataset, an assertion that the missing observations are missing at random, prior to accounting for the missing data mechanism, is incorrect, assuming the relationship among the relevant variables is the same for all susceptible subjects. Furthermore, the use of a faulty MAR assumption in the evaluation of this dependent variable risks serious error in the calculation of the posterior odds and in any conclusion drawn from them. In order to perform a proper evaluation of the opinion on smoking in the workplace in Toronto in early 1993 via the dependent variable of interest in this survey, it is necessary to account for the non-response mechanism in the model structure.

In this analysis, only one simple piece of information, the number of calls, was utilized. A more complete treatment could have been made, had more information been available. Knowledge of the exact number of calls to the non-respondents, instead of a minimum, and the time of day of the calls could have enabled this analysis to be more precise. In addition, knowledge of the type of non-response, refusal or non-availability, and the number of times the non-respondents were actually contacted could have allowed for better classification of the non-respondents. Groves and Couper (1998) point out that statistical errors arising from non-availability and those arising from refusals are likely to differ. As they further comment, the evaluation of how efforts to seek cooperation effect measurement error is an important area of research.

The results illustrated above apply only to this one dependent variable assessing smoking in the workplace in this one dataset. Given the perception that smoking has become less socially acceptable over recent years, it would be reasonable to think that non-response error due to questions about smoking may be more severe than other topics. A comparison of non-response bias including various smoking related questions and others which do not concern smoking may be found in Biemer (2001); this comparison lends no credence to the idea that non-response error is unique to questions relating to smoking.

Although the above results make no implications about the missing data mechanisms in other surveys, there is a clear demonstration here that blindly assuming that the respondents of a survey constitute a random subsample of the population for the variables of interest can be an unwise choice. Information, available at the time of data collection, can enable the evaluation of whether or not the mechanism which causes the non-response is ignorable. In light of this observation, then, it should be of interest to those who work with such data to make use of the available information pertaining to the non-response in the evaluation of that data and to make such information available to others who utilize the dataset. As a general matter, we believe that the collection and analysis of data on where and how respondents were found, as well as how difficult they were to find, is an important future direction for survey methodology and practice.

ACKNOWLEDGEMENTS

This research was funded by National Science Foundation Grant DMS-9801401. The authors thank Shelley Bull for her many helpful comments and suggestions and for assistance in the acquisition of the data and John Eltinge and the anonymous referees and Associate Editor for their valuable comments.

Data from the Attitudes Toward Smoking Legislation, which was funded by Health and Welfare Canada, were made available by the Institute for Social Research at York

University. The data were collected by the Institute for Social Research for Dr. Linda Pederson of the University of Western Ontario, Dr. Shelley Bull of the University of Toronto and Dr. Mary Jane Ashley of the University of Toronto. The principal investigators, the Ontario Ministry of Health and the Institute for Social Research bear no responsibility for the analyses and interpretations presented here.

A. Post-stratification Weighting

HHW_i is the household weight of subject i as described in Northrup (1993).

- Let m = the number of respondents.
- Let r = the cumulative number of adults in the responding households.
- Let h_i = the number of adults in subject i 's household.
- $HHW_i = h_i \cdot m/r$.

Proportions in the sample falling into the following age groups were calculated for both male and female respondents: 18-24 years, 25-44 years, 45-64 years, and over 65 years old. These proportions were then compared to the age/sex distribution in Metropolitan Toronto.

- Let p_{1i} = the proportion of adult Metropolitan Toronto residents falling into the same age/sex category as subject i , as per the 1991 Census.
- Let p_{2i} = the proportion of survey respondents with the same age and sex categories as subject i .
- $W_i = HHW_i \cdot p_{1i}/p_{2i}$, where W_i is the final post-stratification weight used in the analysis.

B. MCMC Implementation

The full MCMC simulation for the NI model consists of a Metropolis algorithm supplemented with the data augmentation described in section 5.3. The following is an overview of the MCMC algorithm. Variables used below are defined in section 5. At each iteration t ,

1. Draw ρ_t for $Beta(s_{t-1} + 1, 2398 - s_{t-1} + 1)$.
2. Impute s_t from $Binomial(\rho_t) \geq 1,429$.
3. Impute $C_{mis_i}^*$: draw $(s_t - 1,429)$ v_i 's from $Geometric(\pi_{t-1})$ and $\forall c_i \in C_{mis}^*, c_i = v_i + 12$.
4. Draw π_t from $Beta(s_t + 1, \sum c_{sus_i} - s_t + 1)$.
5. Impute values for the rest of X_{mis} by utilizing the relationships with the number of calls, as described in section 5.3

6. Update the additional parameters used in the data augmentation of X_{mis} .
 - Update linear regression parameters, β_r and σ_r by drawing directly from the closed form of their posteriors.
 - Update logistic regression parameters, β_l using a Metropolis step on each.
7. Impute Y_{mis} ; $\forall y_i \in y_{\text{mis}}$ draw y_i from a *Multinomial* ($p_0(x_i), p_1(x_i), p_2(x_i)$).
8. Update each β_{kj} using a Metropolis step on the conditional likelihood and a Normal jump function.

REFERENCES

- BIEMER, P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17, 2, 295-320.
- BULL, S. (1994). *Case Studies in Biometry*. Analysis of Attitudes Toward Workplace Smoking Restrictions, chapter 16, New York: Wiley and Sons, 249-270.
- COWLES, M.K., and CARLIN, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, B* 39, 1-38.
- DREW, J.H., and FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the section on Survey Research Methods, American Statistical Association*, Alexandria, VA, 623-628.
- ÉLTINGE, J.L., and YANSANEH, I.S. (1997). Diagnosis for formation of nonresponse adjustment cells, with and application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.
- GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B. (1998). *Bayesian Data Analysis*. Chapter 14, Generalized Linear Models. London: Chapman & Hall.
- GROVES, R.M., and COUPER, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley and Sons.
- HIEDELBERGER, P., and WELCH, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley and Sons.
- MacEACHERN, S.N., and BERLINER, L.M. (1994). Subsampling the Gibbs Sampler. *The American Statistician*, 48, 188-189.
- MALLER, R., and ZHOU, X. (1996). *Survival Analysis with Long Term Survivors*. Chichester, UK: Wiley and Sons.
- NATARAJAN, R., and KASS, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227-237.
- NORTHROP, D.A. (1993). Attitudes Towards Workplace Smoking Legislation: A Survey of Residents of Metropolitan Toronto, Phase III, 1992/93 Technical Documentation. Tech. Rep. Institute for Social Research, York University, Unpublished.
- PEDERSON, L.L., BULL, S.B. and ASHLEY, M.J. (1996). Smoking in the workplace: Do smoking patterns and attitudes reflect the legislative environment? *Tobacco Control*, 5, 39-45.
- PEDERSON, L.L., BULL, S.B., ASHLEY, M.J. and LEFCOE, N.M. (1989). A population survey on legislative measures to restrict smoking in Ontario: 3. Variables related to attitudes of smokers and nonsmokers. *American Journal of Preventive Medicine*, 5, 313-322.
- POTTOFF, R.F., MANTON, K.G. and WOODBURY, M.A. (1993). Correcting for nonavailability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, 88, 1197-1207.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- SHMUELI, G., MINKA, T.P., KADANE, J.B., BORLE, S. and BOATWRIGHT, P. (2001). Using Computational and Mathematical Methods to Explore a New Distribution: The v-Poisson. Technical Report 740, Department of Statistics Carnegie Mellon University.
- TANNER, M.A. and WONG, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-549.

Double Sampling

M.A. HIDIROGLOU¹

ABSTRACT

The theory of double sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information (x) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimate using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. However, it is not necessary for one of the samples to be nested in the other or selected from the same frame. The case of *non-nested* double sampling is dealt with in passing in the classical works on sampling (Des Raj 1968, Cochran 1977). This method is now used in several national statistical agencies.

This paper consolidates double sampling by presenting it in a unified manner. Several examples of surveys used at Statistics Canada illustrate this unification.

KEY WORDS : Double sampling ; Auxiliary data ; Regression ; Optimal.

1. INTRODUCTION

The theory of double-phase sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information (x) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimation by using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. Two-phase sampling is a powerful and cost-effective technique with a long history. Neyman (1938) was first to propose it. Rao (1973) studied double sampling in the context of stratification and analytic studies. Cochran (1977) presented the basic results of two-phase sampling, including the simplest regression estimators for this type of sampling design. More recent work on the subject includes that of Breidt and Fuller (1993), who developed efficient estimation methods for three-phase sampling computations using auxiliary data. Chaudhuri and Roy (1994) focused on the optimal properties of simpler but well-known regression estimators of two-phase sampling. Hidiroglou and Särndal (1998) proposed estimators based on calibration and regression for two-phase sampling to account for the availability of auxiliary data at both levels of the sampling design.

Estimation for nested and non-nested double sampling has been treated separately in the survey literature. However, it is not necessary for one of the samples to be nested within the other, or even be selected from the same survey frame. This case will be termed *non-nested* double sampling. It has been briefly discussed in such classical

books on sampling such as Des Raj (1968) and Cochran (1977). This method is used in several statistical agencies. For example, at Statistics Canada, the Canadian Survey of Employment, Payrolls and Hours (SEPH) is using this sampling procedure (Rancourt and Hidiroglou 1998). In this survey, two independent samples are drawn from two different frames, which nevertheless represent the same universe. The auxiliary data (x), which includes the number of employees and the total amount of payrolls are obtained from a sample selected from a Canada Customs and Revenue Agency administrative data file. These same variables, together with the variables of interest (y), the number of hours worked by employees and summarised earnings, are collected from a sample drawn from the Statistics Canada Business Register. Another example described by Deville (1999) is the case of a household survey conducted at INSEE.

A single estimator can represent the overall estimation process, and the only difference is with respect to variance estimation. This paper is structured as follows. Part 2 sets out the notation. Part 3 describes how the double sampling procedures can be obtained from a single estimator. In Part 4, the estimated variance for the nested and non-nested calibration estimator is presented. Several practical examples are provided in Part 5. Finally, Part 6 contains a brief summary.

2. NOTATION

2.1 Nested Case

The population is represented by $U = \{1, \dots, k, \dots, N\}$. First, a probability sample s_1 ($s_1 \subseteq U$) is selected from population U using a sampling design with inclusion

¹ M.A. Hidiroglou, Business Survey Methods Division, R.H. Coats Building, 11th Floor, Section A, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.
E-mail: hidirogl@statcan.ca.

probability of $\pi_{1k} = P(k \in s_1)$ for the k -th sampled unit in s_1 . Given s_1 , a second sample s_2 ($s_2 \subseteq s_1 \subseteq U$) is drawn from s_1 using a sample design with conditional inclusion probability $\pi_{2k|s_1} = P(k \in s_2 | s_1)$ for the k -th sampled unit in s_2 . Note that the probabilities are conditional since it is assumed that s_1 is known. Figure 1 displays an example of nested sampling.

We assume that $\pi_{1k} > 0$ for all values $k \in U$ and that $\pi_{2k|s_1} > 0$ for all values $k \in s_1$. The weight of a sampled unit k will be denoted by $w_{1k} = 1/\pi_{1k}$ for the first-phase sample and $w_{2k} = 1/\pi_{2k|s_1}$ for the second phase sample. The overall sampling weight of a selected second-phase unit, $k \in s_2$, will therefore be $w_k^* = w_{1k} w_{2k}$.

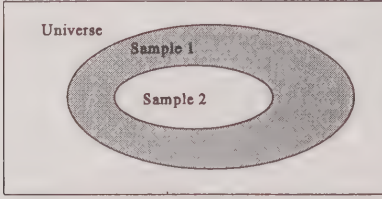


Figure 1. Nested Samples

Let \mathbf{x} denote the auxiliary data vector available with the first-phase sample, and \mathbf{x}_k the value for unit k . We proceed as in Hidirolou and Särndal (1998), that is, we divide \mathbf{x}_k into two parts \mathbf{x}_{1k} and \mathbf{x}_{2k} . The values of the data vector \mathbf{x}_{1k} as assumed to be known for the entire population U , while the values of data vector \mathbf{x}_{2k} are only known for the first-phase sample s_1 .

2.2 Non-nested Case

It is possible for the two samples to be drawn independently from the same frame or even from different (but equivalent) frames. Figures 2 and 3 provide examples of these non-nested cases.

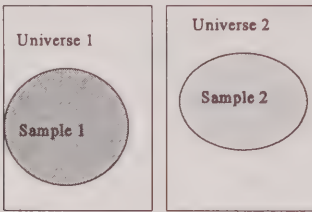


Figure 2. Two independent samples selected from different sample frames

The non-nested case represented by Figure 3 is not considered in this paper. This case can be complicated for arbitrary sampling plans because it is necessary to compute joint inclusion probabilities between the two samples s_1 and s_2 . This computation is simpler when the two samples s_1 and s_2 have been selected using a simple sampling design such as simple random sampling (with or without replacement). It is then possible to use Tam's results (1984)

to obtain the required joint selection probabilities for the computation of the estimated variance for a given estimator of the total $Y = \sum_U y_k$.

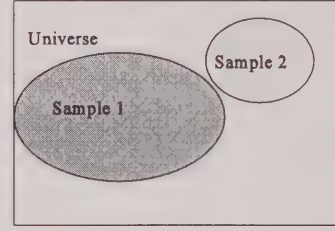


Figure 3. Two samples drawn independently from the same sample frame

For the case that we will study, we assume that samples s_1 and s_2 are drawn independently from two different frames $U_1 = \{1, \dots, k, \dots, N_1\}$ and $U_2 = \{1, \dots, k, \dots, N_2\}$ (see Figure 2). The inclusion probabilities of a sampled unit k are respectively $\pi_{1k}^{(1)} = P(k \in s_1) > 0$ and $\pi_{2k}^{(2)} = P(k \in s_2) > 0$ for samples s_1 ($s_1 \subseteq U_1$) and s_2 ($s_2 \subseteq U_2$). The weight of unit k is $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ for the first sample s_1 and $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ for the second sample s_2 . The superscripts (1) and (2) are used to differentiate between the selection probabilities of the samples drawn in the nested case. The sampling units may differ between the two frames, but these frames represent the same coverage. Examples of such sampling procedures were mentioned in the introduction and more details are provided in the second example given in section 5'3.

Let $\mathbf{x}_k = (\mathbf{x}_{1k}', \mathbf{x}_{2k}')'$, be an auxiliary data vector. We assume that $\mathbf{x}_{1k}^{(1)}$ is known for all units belonging to frame U_1 , while $\mathbf{x}_{1k}^{(1)}$ is only known for sample s_1 . We collect $y_k^{(2)}, \mathbf{x}_{2k}^{(2)}$ from sample s_2 . The \mathbf{x} data collected for corresponding units in samples s_1 and s_2 may differ. The degree in difference between the data values will vary according to the complexity of the sampling unit, and how much these units differ in concept between the two sampling frames. For « simpler » units the data reported for « similar » units in s_1 and s_2 should be equal or almost equal. Departures in the data similarity for the same units in s_1 and s_2 would most likely be due to the different questionnaire wording or due to different respondents filling in the questionnaires. Nevertheless, we assume that $\mathbf{X}_1 = \sum_{U_1} \mathbf{x}_{1k}^{(1)} = \sum_{U_2} \mathbf{x}_{1k}^{(2)}$ since U_1 and U_2 have the same coverage.

3. OPTIMAL ESTIMATOR FOR NESTED AND NON-NESTED SAMPLES

In both cases, nested and non-nested, the objective is to estimate the population total $Y = \sum_U y_k$ where y_k represents the value of unit $k \in U$. An unbiased estimator of Y is $\hat{Y}_{HT} = \sum_{s_2} w_k^* y_k$, where $w_k^* = w_{1k} w_{2k}$ for the nested case and $w_k^* = w_{2k}^{(2)}$ for the non-nested case.

The sampling weight of a unit is modified by multiplying it by the calibration factor obtained using the various levels of the auxiliary data (universe, first-phase sample). The product is called a "calibration weight". Table 1 summarises the available data for the nested and non-nested cases, corresponding to Figures 1 and 2.

Table 1

Data Available for the Population and Samples

Set of Elements	Nested Case	Non-nested Case
Population	\mathbf{x}_{1k} : known for $k \in U$	$\mathbf{x}_{1k}^{(1)}$: known for $k \in U_1$
First sample	\mathbf{x}_k : observed for $k \in s_1$	$\mathbf{x}_k^{(1)}$: observed for $k \in s_{11}$
Second sample	y_k, \mathbf{x}_k : observed for $k \in s_2$	$y_k^{(2)}, \mathbf{x}_k^{(2)}$: observed for $k \in s_{22}$

The following regression estimator is used to estimate the population total Y for nested and non-nested samples:

$$\hat{Y}_{\text{REG}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \mathbf{B}_1 + (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})' \mathbf{B}. \quad (3.1)$$

The various totals corresponding to the auxiliary data \mathbf{x} and y -variable of interest given in equation (3.1) are provided in Table 2.

It is assumed that the variances, $V(\hat{Y}_{\text{HT}})$, and covariances $\text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{X}}')$, $\text{Cov}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_1')$, $\text{Cov}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}')$, $\text{Cov}(\hat{Y}_{\text{HT}}, \hat{\mathbf{X}}')$ and $\text{Cov}(\hat{Y}_{\text{HT}}, \hat{\mathbf{X}}_1')$, are known or estimable.

To simplify the notation, we drop the superscripts for the remainder of this section. The estimation of the parameters, \mathbf{B} and \mathbf{B}_1 as well as of their associated variance, reflect that we have sampled differently for the nested and non-nested cases. The estimators of \mathbf{B} and \mathbf{B}_1 are obtained by minimising the variance of \hat{Y}_{REG} . This variance is:

$$\begin{aligned} V(\hat{Y}_{\text{REG}}) &= V(\hat{Y}_{\text{HT}}) + \mathbf{B}_1' V(\hat{\mathbf{X}}_1) \mathbf{B}_1 + \mathbf{B}' V(\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}) \mathbf{B} \\ &\quad - 2 \text{Cov}(\hat{Y}_{\text{HT}}, \hat{\mathbf{X}}_1') \mathbf{B}_1 + 2 \text{Cov}(\hat{Y}_{\text{HT}}, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})') \mathbf{B} \\ &\quad - 2 \mathbf{B}_1' \text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})') \mathbf{B}. \end{aligned} \quad (3.2)$$

Deriving (3.2) with respect to \mathbf{B} and \mathbf{B}_1 , we obtain the following two equations:

$$\begin{aligned} V(\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}) \mathbf{B} + \text{Cov}((\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}), \hat{Y}_{\text{HT}}) \\ - \text{Cov}((\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}), \hat{\mathbf{X}}_1') \mathbf{B}_1 = \mathbf{0} \end{aligned} \quad (3.3)$$

and

$$-\text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})') \mathbf{B} - \text{Cov}(\hat{\mathbf{X}}_1, \hat{Y}_{\text{HT}}) + V(\hat{\mathbf{X}}_1) \mathbf{B}_1 = \mathbf{0}. \quad (3.4)$$

Solving the system of equations (3.3) and (3.4), we obtain the required parameters \mathbf{B} and \mathbf{B}_1 . That is:

$$\mathbf{B} = \mathbf{T}^{-1} \mathbf{H} \quad (3.5)$$

where

$$\begin{aligned} \mathbf{T} &= V(\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}) - (\text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})'))' \\ &\quad V^{-1}(\hat{\mathbf{X}}_1) (\text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})')), \end{aligned}$$

$$\begin{aligned} \mathbf{H} &= (\text{Cov}((\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}}), \hat{Y}_{\text{HT}})) \\ &\quad + (\text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})'))' V^{-1}(\hat{\mathbf{X}}_1) \text{Cov}(\hat{\mathbf{X}}_1, \hat{Y}_{\text{HT}}) \end{aligned}$$

and

$$\mathbf{B}_1 = \mathbf{T}_1^{-1} \mathbf{H}_1 \quad (3.6)$$

where

$$\mathbf{T}_1 = V(\hat{\mathbf{X}}_1),$$

and

$$\mathbf{H}_1 = \text{Cov}(\hat{\mathbf{X}}_1, \hat{Y}_{\text{HT}}) + \text{Cov}(\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})')' \mathbf{B}.$$

Table 2

Sums of the Auxiliary Data \mathbf{x} and y for Nested and Non-nested Cases

Set of Elements	Nested Case	Non-nested Case
Population	$\mathbf{X}_1 = \sum_U \mathbf{x}_{1k}$	$\mathbf{X}_1 = \sum_{U_1} \mathbf{x}_{1k}^{(1)}$
First sample	$\hat{\mathbf{X}}_1 = \sum_{s_1} w_{1k} \mathbf{x}_{1k}; \hat{\mathbf{X}} = \sum_{s_1} w_{1k} \mathbf{x}_k$	$\hat{\mathbf{X}}_1 = \sum_{s_{11}} w_{1k} \mathbf{x}_{1k}^{(1)}; \hat{\mathbf{X}} = \sum_{s_{11}} w_{1k} \mathbf{x}_k^{(1)}$
Second sample	$\hat{\hat{\mathbf{X}}}_1 = \sum_{s_2} w_{2k} \mathbf{x}_{1k}; \hat{\hat{\mathbf{X}}} = \sum_{s_2} w_{2k} \mathbf{x}_k$ $\hat{Y}_{\text{HT}} = \sum_{s_2} w_{2k} y_k$	$\hat{\hat{\mathbf{X}}}_1 = \sum_{s_{22}} w_{2k} \mathbf{x}_{1k}^{(2)}; \hat{\hat{\mathbf{X}}} = \sum_{s_{22}} w_{2k} \mathbf{x}_k^{(2)}$ $\hat{Y}_{\text{HT}} = \sum_{s_{22}} w_{2k} y_k^{(2)}$

Result 1: An optimal regression estimator for the nested and non-nested samples is:

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \hat{\mathbf{B}}_{1,\text{OPT}} + (\hat{\mathbf{X}} - \hat{\mathbf{X}})' \hat{\mathbf{B}}_{\text{OPT}} \quad (3.7)$$

where

$$\hat{\mathbf{B}}_{\text{OPT}} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{H}} \quad (3.8)$$

and

$$\hat{\mathbf{B}}_{1,\text{OPT}} = \hat{\mathbf{T}}_1^{-1} \hat{\mathbf{H}}_1. \quad (3.9)$$

$\hat{\mathbf{T}}_1$, $\hat{\mathbf{H}}_1$, $\hat{\mathbf{T}}$ and $\hat{\mathbf{H}}$ are the estimated values of \mathbf{T}_1 , \mathbf{H}_1 , \mathbf{T} and \mathbf{H} , and they are obtained using a framework leading to the inference based on the sampling design. These values are dependent on the sample selection scheme. The population variance of \hat{Y}_{OPT} and its associated estimated variance depend on whether or not the samples are nested or non-nested. Since the regression vectors are optimal, it follows that the regression estimator \hat{Y}_{OPT} is also optimal. The optimal form has been discussed by Montanari (1987, 1998, and 2000) for the case of a single phase sampling design.

3.1 The Case of Nested Double Sampling

The theory for this case is developed using a conditional approach. Suppose that two parameters are given by θ_1 and θ_2 , and that they are estimated by $\hat{\theta}_1$ and $\hat{\theta}_2$ from sample s_2 . If we condition on the realised sample s_1 , then the following well-known results hold:

- (i) The expectation of $\hat{\theta}$ is $E(\hat{\theta}) = E_1 E_2(\hat{\theta} | s_1)$, where E_2 denotes the expectation of $\hat{\theta}$ given s_1 .
- (ii) The variance of $\hat{\theta}$ is

$$V(\hat{\theta}) = E_1 V_2(\hat{\theta} | s_1) + V_1 E_2(\hat{\theta} | s_1). \quad (3.10)$$

- (iii) The covariance between $\hat{\theta}_1$ and $\hat{\theta}_2$ is:

$$\begin{aligned} \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) &= E_1 \text{Cov}_2((\hat{\theta}_1, \hat{\theta}_2)' | s_1) \\ &+ \text{Cov}_1(E_2(\hat{\theta}_1 | s_1), E_2(\hat{\theta}_2 | s_1)). \end{aligned}$$

The various components of $\hat{\mathbf{T}}$, $\hat{\mathbf{H}}$, $\hat{\mathbf{T}}_1$ and of $\hat{\mathbf{H}}_1$ will be estimated assuming an arbitrary sampling design with a non-fixed sample size. The case of a fixed size sampling design follows easily as it is a special case of the arbitrary sampling design. Using expressions (i) – (iii), we can express the terms defining parameter \mathbf{B} as:

$$\text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{X}}') = \text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{X}}') = V(\hat{\mathbf{X}});$$

$$\text{Cov}(\hat{\mathbf{Y}}_{\text{HT}}, \hat{\mathbf{X}}') = \text{Cov}(\hat{\mathbf{Y}}_{\text{HT}}, \hat{\mathbf{X}}');$$

$$V(\hat{\mathbf{X}} - \hat{\mathbf{X}}) = E_1 \left[\sum \sum_{s_1} c_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_\ell' \right];$$

$$\text{Cov}[\hat{\mathbf{X}}_1, (\hat{\mathbf{X}} - \hat{\mathbf{X}})'] = \mathbf{0};$$

and

$$\text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}_{\text{HT}}) = \text{Cov}(\hat{\mathbf{X}}, \hat{\mathbf{Y}}_{\text{HT}}) + E_1 \left[\sum \sum_{s_1} c_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_\ell' \right]; \quad (3.11)$$

where $c_{2k\ell | s_1} = (\pi_{2k\ell | s_1} - \pi_{2k | s_1} \pi_{2\ell | s_1}) / \pi_{1k}^* \pi_{1\ell}^*$ and $\hat{\mathbf{Y}}_{\text{HT}} = \sum_{s_1} y_k / \pi_{1k}$. The inclusion probabilities in these expressions are $\pi_{2k\ell | s_1} = \Pr(k, \ell \in s_2 | s_1)$ and $\pi_{1k}^* = \pi_{1k} \pi_{2k | s_1}$. We can express \mathbf{B} more simply as:

$$\mathbf{B} = \left[E_1 \left(\sum \sum_{s_1} c_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_\ell' \right) \right]^{-1} E_1 \left[\sum \sum_{s_1} c_{2k\ell | s_1} \mathbf{x}_k y_\ell \right] \quad (3.12)$$

and the corresponding optimal estimator is given by:

$$\begin{aligned} \hat{\mathbf{B}}_{\text{OPT}} &= \left[\sum \sum_{s_2} \hat{c}_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_\ell' \right]^{-1} \\ &\left[\sum \sum_{s_2} \hat{c}_{2k\ell | s_1} \mathbf{x}_k y_\ell \right] \end{aligned} \quad (3.13)$$

where $\hat{c}_{2k\ell | s_1} = c_{2k\ell | s_1} / \pi_{2k\ell | s_1}$.

The optimal regression estimator $\hat{\mathbf{B}}_{1,\text{OPT}}$ is given by (3.9) with

$$\hat{\mathbf{T}}_1 = \hat{\mathbf{V}}(\hat{\mathbf{X}}_1)$$

and

$$\begin{aligned} \hat{\mathbf{H}}_1 &= \text{C}\hat{\text{O}}\text{V}(\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_{\text{HT}}) + \text{C}\hat{\text{O}}\text{V}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}') \hat{\mathbf{B}}_{\text{OPT}} \\ &- \text{C}\hat{\text{O}}\text{V}(\hat{\mathbf{X}}_1, \hat{\mathbf{X}}') \hat{\mathbf{B}}_{\text{OPT}}. \end{aligned}$$

Each component defining $\hat{\mathbf{T}}_1$ and $\hat{\mathbf{H}}_1$ is estimated as follows. We first estimate $V(\hat{\mathbf{X}}_1) = \sum \sum_{s_1} c_{1k\ell} \mathbf{x}_{1k} \mathbf{x}_{1\ell}'$ by

$$\hat{V}(\hat{\mathbf{X}}_1) = \sum \sum_{s_1} \hat{c}_{1k\ell} \mathbf{x}_{1k} \mathbf{x}_{1\ell}' \quad (3.14)$$

where $c_{1k\ell} = (\pi_{1k\ell} - \pi_{1k} \pi_{1\ell}) / (\pi_{1k} \pi_{1\ell})$ and $\hat{c}_{1k\ell} = c_{1k\ell} / \pi_{1k\ell}$.

Next, since

$$\begin{aligned}
\text{Cov}(\hat{X}_1, \hat{Y}_{HT}) &= E_1 \text{Cov}_2 \left[\left(\hat{X}_1, \hat{Y}_{HT} \right) | s_1 \right] \\
&\quad + \text{Cov}_1 \left[E_2(\hat{X}_1 | s_1), E_2(\hat{Y}_{HT} | s_1) \right] \\
&= \text{Cov}_1(\hat{X}_1, \hat{Y}_{HT}) \\
&= \sum \sum_{U_1} c_{1k\ell} x_{1k} y_{1\ell} \quad (3.15)
\end{aligned}$$

we estimate $\text{Cov}(\hat{X}_1, \hat{Y}_{HT})$ by

$$\widehat{\text{Cov}}(\hat{X}_1, \hat{Y}_{HT}) = \sum \sum_{s_2} c_{1k\ell}^* x_{1k} y_{1\ell} \quad (3.16)$$

where

$$\begin{aligned}
c_{1k\ell}^* &= c_{1k\ell} / \pi_{k\ell}^*, \quad \pi_{k\ell}^* = \pi_{1k\ell} \pi_{2k\ell} | s_1, \\
\pi_{1k\ell} &= \Pr(k, \ell \in s_1), \\
\pi_{2k\ell} | s_1 &= \Pr(k, \ell \in s_2 | s_1) \\
\text{and } \pi_k^* &= \pi_{1k} \pi_{2k} | s_1.
\end{aligned}$$

Similarly,

$$\widehat{\text{Cov}}(\hat{X}_1, \hat{X}') = \sum \sum_{s_2} c_{1k\ell}^* x_{1k} x'_{\ell} \quad (3.17)$$

and

$$\widehat{\text{Cov}}(\hat{X}_1, \hat{X}') = \sum \sum_{s_1} \hat{c}_{1k\ell} x_{1k} x'_{\ell}. \quad (3.18)$$

Hence, in the case of nested double sampling the optimal estimator of B_1 is given by:

$$\begin{aligned}
\hat{B}_{1, \text{OPT}} &= \left(\hat{V}(\hat{X}_1) \right)^{-1} \left[\widehat{\text{Cov}}(\hat{X}_1, \hat{Y}_{HT}) \right. \\
&\quad \left. + \left(\widehat{\text{Cov}}(\hat{X}_1, \hat{X}') - \widehat{\text{Cov}}(\hat{X}_1, \hat{X}) \right) \hat{B}_{\text{OPT}} \right] \quad (3.19)
\end{aligned}$$

where the components of $\hat{B}_{1, \text{OPT}}$ have been defined by expressions (3.14) – (3.18).

The optimal form of estimators $\hat{B}_{1, \text{OPT}}$ and \hat{B}_{OPT} has its advantages and disadvantages. One of the biggest advantages of the optimal form, as reported by Cassady and Valliant (1993), Rao (1994), and Montanari (2000), is that it has good conditional inference properties (by conditioning on the auxiliary variable x). As Montanari (2000) observed, the asymptotic optimality of \hat{Y}_{OPT} is strictly a property based on the sampling design and achieved conditionally on the finite population. The biggest disadvantage of the optimal estimator is that it requires the computation of joint inclusion probabilities.

We can, however, use the optimal form, and express it more simply for several sampling designs. For sampling designs where the sample selection is with unequal probability and without replacement, we can bypass the computation of the joint probability by approximating the exact variance. Several authors, including Hartley and Rao

(1962), Deville (1999), Berger (1998), Rósen (2000) and Brewer (2000) proposed such approximating procedures. Recently, Tillé (2001) proposed the following approximation for the estimated variance of $\hat{Y}_{HT} = \sum_s y_k / \pi_k$ in the context of single-phase sampling, where

$$\begin{aligned}
\hat{V}(\hat{Y}_{HT}) &= \sum_s \frac{c_k}{\pi_k^2} (y_k - y_k^*)^2 \\
&= \sum_s c_k \left(\frac{y_k}{\pi_k} - \tilde{y} \right)^2. \quad (3.20)
\end{aligned}$$

Here, c_k is the variable used as the approximation, $y_k^* = \pi_k \sum_s c_\ell y_\ell / \pi_\ell / \sum_s c_\ell$, $\tilde{y} = y_k^* / \pi_k$, and π_k is the probability of selection of a given unit k . Tillé (2001) provided several examples of the c_k values for various sampling schemes.

This formula is exact in the case of a stratified simple sampling design drawn without replacement in each stratum U_h ($h = 1, \dots, L$) of population U . Let k be a sampled unit in sample s_h from stratum U_h , then $c_k = n_h / (n_h - 1) (1 - n_h / N_h)$ if $k \in U_h$ and 0 otherwise, and $\pi_k = n_h / N_h$ if $k \in U_h$ and 0 otherwise. This gives us the exact estimated variance, $\hat{V}(\hat{Y}_{HT}) = \sum_{h=1}^L N_h^2 (1 - n_h / N_h) \sum_{s_h} (y_k - \bar{y}_h)^2 / n_h (n_h - 1)$. The formula is also exact in the case of a stratified sampling design where the sample is selected with replacement. Here $c_k = 1$ for all units belonging to stratum U_h and zero otherwise. Using this approximation, the double sums appearing in \hat{B}_{OPT} and $\hat{B}_{1, \text{OPT}}$ can be expressed as simple sums. Hidiroglou and Särndal (1998) bypassed the problem of double sums in estimating B and B_1 by proposing the GREG estimator, \hat{Y}_{GREG} , for a nested two-phase sampling design. Their estimator is given by:

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{HT} + (X_1 - \hat{X}_1)' \hat{B}_{1, \text{GREG}} + (\hat{X} - \hat{X}')' \hat{B}_{\text{GREG}}$$

where

$$\hat{B}_{\text{GREG}} = \left(\sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} x_k x'_k}{\sigma_{2k}^2} \right)^{-1} \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} x_k y_k}{\sigma_{2k}^2}, \quad (3.21)$$

$$\begin{aligned}
\hat{B}_{1, \text{GREG}} &= \left(\sum_{s_1} \frac{w_{1k} x_{1k} x'_{1k}}{\sigma_{1k}^2} \right)^{-1} \\
&\quad \left\{ \sum_{s_2} \frac{w_k^* x_{1k} y_k}{\sigma_{1k}^2} + \sum_{s_1} \frac{w_{1k} x_{1k} x'_k}{\sigma_{1k}^2} \hat{B}_{\text{GREG}} \right. \\
&\quad \left. - \sum_{s_2} \frac{w_k^* x_{1k} x'_k}{\sigma_{1k}^2} \hat{B}_{\text{GREG}} \right\} \quad (3.22)
\end{aligned}$$

with $\{\sigma_{1k}^2 : k \in s_1\}$ and $\{\sigma_{2k}^2 : k \in s_2\}$ being predetermined positive factors.

Estimators \hat{B}_{GREG} and $\hat{B}_{1, \text{GREG}}$ can be justified either by assuming different regression models for each phase or by using two successive calibrations. For the calibration approach, calibration weights \tilde{w}_{1k} associated with the first-phase are first obtained, and they satisfy the calibration equation $\sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. These calibration weights can be expressed as the product of sample weights w_{1k} and a calibration factor g_{1k} where:

$$g_{1k} = 1 + \left(\sum_U \mathbf{x}_{1k} - \sum_{s_1} w_{1k} \mathbf{x}_{1k} \right)' \left(\sum_{s_1} w_{1k} \frac{\mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2} \right)^{-1} \frac{\mathbf{x}_{1k}}{\sigma_{1k}^2} \quad (3.23)$$

for $k \in s_1$.

The first-phase calibration weights \tilde{w}_{1k} are then used as initial weights to compute the overall calibration weights \tilde{w}_k^* . These overall calibration weights satisfy the second-phase calibration equation $\sum_{s_2} \tilde{w}_{1k}^* \mathbf{x}_k = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_k$. The estimator of the total, \hat{Y}_{GREG} , can be expressed as the sum of the product of the overall calibration weight \tilde{w}_k^* and the associated y -value, that is $\hat{Y}_{\text{GREG}} = \sum_{s_2} \tilde{w}_k^* y_k$. The calibrated overall weights can be expressed as $\tilde{w}_k^* = w_k^* g_k^*$, where $g_k^* = g_{1k} g_{2k}$. Here, g_{1k} is given by (3.23), while g_{2k} is equal to

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k} \mathbf{x}_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} \mathbf{x}_k \right)' \left(\sum_{s_1} \frac{\tilde{w}_{1k} w_{2k} \mathbf{x}_k \mathbf{x}_k'}{\sigma_{2k}^2} \right)^{-1} \frac{\mathbf{x}_k}{\sigma_{2k}^2} \quad (3.24)$$

for $k \in s_2$.

Comment: The estimators of $\hat{B}_{1, \text{GREG}}$ (3.21) and \hat{B}_{GREG} (3.22) correspond to Hidioglou and Särndal's (1998) *additive case* and have the same form as the optimal regression estimators $\hat{B}_{1, \text{OPT}}$ (3.8) and \hat{B}_{OPT} (3.9). Indeed, the components of the estimator of B are obtained by respectively estimating T by $(\sum_{s_2} w_{1k} w_{2k} \mathbf{x}_k \mathbf{x}_k' / \sigma_{2k}^2)$ and H by $\sum_{s_2} w_{1k} w_{2k} \mathbf{x}_k y_k / \sigma_{2k}^2$. The second terms of H and T are exactly equal to zero. Similarly, to estimate B_1 , the component T_1 is estimated by $\sum_{s_1} w_{1k} \mathbf{x}_{1k} \mathbf{x}_{1k}' / \sigma_{1k}^2$, while H_1 is estimated by

$$\sum_{s_2} \frac{w_k^* \mathbf{x}_{1k} y_{2k}}{\sigma_{1k}^2} + \left(\sum_{s_1} \frac{w_{1k} \mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2} - \sum_{s_2} \frac{w_k^* \mathbf{x}_{1k} \mathbf{x}_{1k}'}{\sigma_{1k}^2} \right)' \hat{B}_{\text{GREG}}$$

The estimated variance of $\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \hat{B}_{1, \text{GREG}} + (\hat{\mathbf{X}} - \hat{\mathbf{X}})' \hat{B}_{\text{GREG}}$ is presented in Hidioglou and Särndal (1998).

Comment: The efficiency of the GREG, as stated in Särndal, Swensson and Wretman (1992), requires that the proposed model be correct. Furthermore, if the sample size is large enough, optimal estimators are more efficient (Rao 1994) than the GREG. However, if the sample size is

relatively small, one disadvantage of the optimal form OPT is that it is generally less stable and more complex to compute than the GREG. Furthermore, an additional consequence of a relatively small sample size, as reported by Särndal (1996), and illustrated by simulation by Montanari (2000), is that if the sample size is relatively small, then the optimal form is not significantly more efficient than the GREG. It is even possible for the estimated variance to be greater than that associated with the GREG.

3.2 The Case of Non-nested Double Sampling

Deville (1999) considered the non-nested case (Figure 2) by assuming that \mathbf{x}_{2k} is known for s_1 and s_2 . The optimal regression estimator is:

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\hat{\mathbf{X}}_2 - \hat{\mathbf{X}}_2)' \hat{B}_{2, \text{OPT}} \quad (3.25)$$

where $\hat{Y}_{\text{HT}} = \sum_{s_2} w_{2k} y_k$, $\hat{\mathbf{X}}_2 = \sum_{s_1} w_{1k} \mathbf{x}_{2k}$, $\hat{\mathbf{X}}_2 = \sum_{s_2} w_{2k} \mathbf{x}_{2k}$. The optimal estimator for $B_2 = (\sum_{U_2} \mathbf{x}_{2k} \mathbf{x}_{2k}')^{-1} \sum_{U_2} \mathbf{x}_{2k} y_k$ is $\hat{B}_{2, \text{OPT}} = (\hat{V}(\hat{\mathbf{X}}_2) + \hat{V}(\hat{\mathbf{X}}_2))^{-1} \text{Cov}(\hat{Y}_{\text{HT}}, \hat{\mathbf{X}}_2')$ if the two sampling frames U_1 and U_2 are independent. The form of the variance and of the covariance terms defining $\hat{B}_{2, \text{OPT}}$ depends on the sampling design of s_1 and s_2 .

The accuracy of the estimator of X_2 can be improved by minimising the variance of $\tilde{X}_2 = A_2 \hat{X}_2 + (I - A_2) \hat{\tilde{X}}_2$ yielding, $A_2 = (V(\hat{\tilde{X}}_2) + V(\hat{\tilde{X}}_2))^{-1} V(\hat{\tilde{X}}_2)$. Assuming that $V(\hat{\tilde{X}}_2)$ is approximately a multiple of $V(\hat{\tilde{X}}_2)$, that is $V(\hat{\tilde{X}}_2) \doteq \alpha_2 V(\hat{\tilde{X}}_2)$, we obtain $A_2 \doteq I / (1 + \alpha_2)$ where I is the identity matrix has the same dimension as the covariance matrix $V(\hat{\tilde{X}}_2)$. The optimal value of α_2 is obtained by minimising the variance of \tilde{X}_2 . A sub-optimal but adequate choice, suggested by Deville (1999), for α_2 is $\alpha_2 = n_1 / (n_1 + n_2)$, where n_1 and n_2 are the respective sizes of samples s_1 and s_2 . Note that Korn and Graubart (1999) also made the same suggestion in the context of combining two totals estimated from two different sources. Substituting \tilde{X}_2 in place of \hat{X}_2 in expression (3.25), yields

$$\tilde{X}_2 - \hat{\tilde{X}}_2 = (\hat{\tilde{X}}_2 - \hat{\tilde{X}}_2) / (1 + \alpha_2). \quad (3.26)$$

The estimator of the population total Y , is:

$$\tilde{\tilde{Y}}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\tilde{\tilde{X}}_2 - \hat{\tilde{X}}_2)' \tilde{\tilde{B}}_{2, \text{OPT}} \quad (3.27)$$

where

$$\tilde{\tilde{B}}_{2, \text{OPT}} = - \left[\hat{V}(\tilde{\tilde{X}}_2 - \hat{\tilde{X}}_2) \right]^{-1} \text{Cov}(\hat{Y}_{\text{HT}}, (\tilde{\tilde{X}}_2 - \hat{\tilde{X}}_2)'). \quad (3.28)$$

If (3.26) is substituted in (3.28), we can re-express $\tilde{\tilde{B}}_{2, \text{OPT}}$ as:

$$\tilde{\tilde{B}}_{2, \text{OPT}} = \left[\hat{V}(\hat{\tilde{X}}_2) \right]^{-1} \text{Cov}(\hat{Y}_{\text{HT}}, \hat{\tilde{X}}_2). \quad (3.29)$$

Comment: We see that \hat{Y}_{OPT} (3.25) is exactly equal to \tilde{Y}_{OPT} (3.27). This implies that there was no advantage in using a better estimator of X_2 to estimate Y . However, the estimator $\tilde{B}_{2,\text{OPT}}$ associated with \tilde{Y}_{OPT} looks more like a traditional regression estimator than the regression estimator $\hat{B}_{2,\text{OPT}}$ associated with \hat{Y}_{OPT} .

Note that the GREG estimator for the case where \tilde{X}_2 is used instead of \hat{X}_2 is:

$$\tilde{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\tilde{X}_2 - \hat{X}_2)' \tilde{B}_{2,\text{GREG}} \quad (3.30)$$

where

$$\tilde{B}_{2,\text{GREG}} = \left(\sum_{s_2} w_{2k} \mathbf{x}_k^{(2)} \mathbf{x}_k'^{(2)} / \sigma_{2k}^2 \right)^{-1} \sum_{s_2} w_{2k} \mathbf{x}_k^{(2)} y_k^{(2)} / \sigma_{2k}^2.$$

Furthermore, if we also know $\mathbf{x}_{1k}^{(1)}$ for $k \in U_1$ where $X_1 = \sum_{U_1} \mathbf{x}_{1k}^{(1)}$, we can consider the regression estimator

$$\tilde{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \tilde{\mathbf{X}}_1)' \tilde{B}_{1,\text{OPT}} + (\tilde{X} - \hat{X})' \tilde{B}_{\text{OPT}} \quad (3.31)$$

We obtain \tilde{X} by minimising the linear combination $A\hat{X} + (I - A)\tilde{X}$ and $V(\tilde{X}) = \alpha V(\hat{X})$. The difference between \tilde{X} and \hat{X} can be re-expressed as

$$\tilde{X} - \hat{X} = (\tilde{X} - \hat{X}) / (1 + \alpha). \quad (3.32)$$

Given that s_1 and s_2 are independent samples, it can be shown that:

$$\tilde{B}_{\text{OPT}} = [\hat{V}(\hat{X})]^{-1} \text{C}\hat{\text{Ov}}(\hat{X}, \hat{Y}_{\text{HT}}) \quad (3.33)$$

and that

$$\tilde{B}_{1,\text{OPT}} = [\hat{V}(\hat{X}_1)]^{-1} [\text{C}\hat{\text{Ov}}(\hat{X}_1, \hat{Y}_{\text{HT}})]. \quad (3.34)$$

The components of \tilde{B}_{OPT} are estimated by:

$$\hat{V}(\hat{X}) = \sum \sum_{s_2} \hat{c}_{2k\ell} \mathbf{x}_k^{(2)} \mathbf{x}_\ell'^{(2)} \quad (3.35)$$

and

$$\text{C}\hat{\text{Ov}}(\hat{X}, \hat{Y}_{\text{HT}}) = \sum \sum_{s_2} \hat{c}_{2k\ell} \mathbf{x}_k^{(2)} y_\ell^{(2)} \quad (3.36)$$

whereas the components of $\tilde{B}_{1,\text{OPT}}$ are estimated by:

$$\hat{V}(\hat{X}_1) = \sum \sum_{s_2} \hat{c}_{2k\ell} \mathbf{x}_{1k}^{(2)} \mathbf{x}_{1\ell}'^{(2)} \quad (3.37)$$

and

$$\text{C}\hat{\text{Ov}}(\hat{X}_1, \hat{Y}_{\text{HT}}) = \sum \sum_{s_2} \hat{c}_{2k\ell} \mathbf{x}_{1k}^{(2)} y_\ell^{(2)} \quad (3.38)$$

where

$$\hat{c}_{2k\ell} = \frac{\pi_{2k\ell} - \pi_{2k}\pi_{2\ell}}{(\pi_{2k\ell})(\pi_{2k}\pi_{2\ell})}.$$

Approximation (3.20) can also be used to estimate the terms (3.35) – (3.38). The corresponding GREG which bypasses the computation of joint selection probabilities is given by:

$$\tilde{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \tilde{\mathbf{X}}_1)' \tilde{B}_{1,\text{GREG}} + (\tilde{X} - \hat{X})' \tilde{B}_{\text{GREG}} \quad (3.39)$$

where $\mathbf{X}_1 = \sum_{U_1} \mathbf{x}_{1k}^{(1)}$, $\tilde{\mathbf{X}}_1 = \sum_{s_1} w_{1k} \mathbf{x}_{1k}^{(1)}$, $\hat{X} = \sum_{s_1} w_{1k} \mathbf{x}_k^{(1)}$ and $\tilde{X} = \sum_{s_2} w_{2k} \mathbf{x}_k^{(2)}$.

GREG-type regression estimators in equation (3.39) are estimated by

$$\tilde{B}_{1,\text{GREG}} = \left(\sum_{s_2} w_{2k} \frac{\mathbf{x}_{1k}^{(2)} \mathbf{x}_{1k}'^{(2)}}{\sigma_{1k}^2} \right)^{-1} \sum_{s_2} w_{2k} \frac{\mathbf{x}_{1k}^{(2)} y_k^{(2)}}{\sigma_{1k}^2} \quad (3.40)$$

and

$$\tilde{B}_{\text{GREG}} = \left(\sum_{s_2} w_{2k} \frac{\mathbf{x}_k^{(2)} \mathbf{x}_k'^{(2)}}{\sigma_{2k}^2} \right)^{-1} \sum_{s_2} w_{2k} \frac{\mathbf{x}_k^{(2)} y_k^{(2)}}{\sigma_{2k}^2}. \quad (3.41)$$

4. ESTIMATOR OF THE VARIANCE FOR THE OPTIMAL REGRESSION ESTIMATOR

4.1 Nested Double Sampling

Recall that the optimal regression estimator of Y is given by

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' \hat{B}_{1,\text{OPT}} + (\hat{X} - \hat{X})' \hat{B}_{\text{OPT}} \quad (4.1)$$

To obtain the estimated variance of (4.1), we re-express the terms associated with the y -variable within \hat{B}_{OPT} and $\hat{B}_{1,\text{OPT}}$ as a simple sums instead of double sums. Montanari (1998) described this algebra for an arbitrary single-phase sampling design. Following Montanari (1998), and adapting the single-phase algebra to double sampling, we obtain:

$$\begin{aligned} \hat{B}_{\text{OPT}} &= \left[\sum \sum_{s_2} \hat{c}_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_\ell' \right]^{-1} \left[\sum \sum_{s_2} \hat{c}_{2k\ell | s_1} \mathbf{x}_k y_\ell \right] \\ &= \left[\sum \sum_{s_2} \hat{c}_{2k\ell | s_1} \mathbf{x}_k \mathbf{x}_k' \right]^{-1} \left[\sum_{s_2} \frac{a_{2k}}{\pi_k^*} y_k \right] \end{aligned} \quad (4.2)$$

where

$$a_{2k} = \frac{1 - \pi_{2k | s_1}}{\pi_k^*} \mathbf{x}_k + \sum_{\ell \neq k} \frac{(\pi_{2k\ell | s_1} - \pi_{2k | s_1} \pi_{2\ell | s_1})}{\pi_{2k\ell | s_1} \pi_\ell^*} \mathbf{x}_\ell.$$

We approximate $\hat{B}_{1,\text{OPT}}$ given by (3.15) by $[\hat{V}(\hat{X}_1)]^{-1} [\text{C}\hat{\text{Ov}}(\hat{X}_1, \hat{Y}_{\text{HT}})]$, and hence,

$$\begin{aligned}\hat{B}_{1,OPT} &\doteq [\hat{V}(\hat{X}_1)]^{-1} [\text{Cov}(\hat{X}_1, \hat{Y}_{HT})] \\ &= \left[\sum_{s_1} \sum_{\ell \neq k} \hat{c}_{1k\ell} \mathbf{x}_{1k} \mathbf{x}_{1\ell}' \right]^{-1} \left[\sum_{s_1} \frac{a_{1k}}{\pi_{1k}} y_k \right] \quad (4.3)\end{aligned}$$

where

$$a_{1k} = \frac{1 - \pi_{1k}}{\pi_{1k}} x_{1k} + \sum_{\ell \neq k} \frac{(\pi_{1k\ell} - \pi_{1k} \pi_{1\ell})}{\pi_{1\ell} \pi_{1k\ell}} x_{1\ell}.$$

By substituting (4.2) and (4.3) in (4.1), and by subtracting the population total Y , we get:

$$\begin{aligned}\hat{Y}_{OPT} - Y &\doteq \left(\sum_{s_1} g_{1k} \frac{y_k}{\pi_{1k}} - \sum_U y_k \right) \\ &+ \left(\sum_{s_2} g_{2k} \frac{y_k}{\pi_k} - \sum_{s_1} \frac{y_k}{\pi_{1k}} \right) \quad (4.4)\end{aligned}$$

where

$$g_{1k} = 1 + (\mathbf{X}_1 - \hat{\mathbf{X}}_1)' (\hat{V}(\hat{\mathbf{X}}_1))^{-1} a_{1k} \quad \text{for } k \in s_1 \quad (4.5)$$

and

$$g_{2k} = 1 + (\hat{\mathbf{X}} - \hat{\hat{\mathbf{X}}})' (\hat{V}(\hat{\hat{\mathbf{X}}}))^{-1} a_{2k} \quad \text{for } k \in s_2. \quad (4.6)$$

Result 2: The estimated variance of \hat{Y}_{OPT} defined by equation (4.1) is:

$$\begin{aligned}\hat{V}(\hat{Y}_{OPT}) &= \sum_{s_2} \sum_{\ell} c_{1k\ell}^* g_{1k} g_{1\ell} e_{1k} e_{1\ell} \\ &+ \sum_{s_2} \sum_{\ell} c_{2k\ell}^* g_{2k} g_{2\ell} e_{2k} e_{2\ell} \quad (4.7)\end{aligned}$$

where

$$\begin{aligned}c_{1k\ell}^* &= \frac{(\pi_{1k\ell} - \pi_{1k} \pi_{1\ell})}{\pi_{k\ell}^* \pi_{1k} \pi_{1\ell}}, \\ c_{2k\ell}^* &= \frac{(\pi_{2k\ell|s_1} - \pi_{2k|s_1} \pi_{2\ell|s_1})}{\pi_{2k\ell|s_1} \pi_k^* \pi_{\ell}^*}, \\ e_{1k} &= y_k - \mathbf{x}_{1k}' \hat{\mathbf{B}}_{1,OPT}; \\ e_{2k} &= y_k - \mathbf{x}_k' \hat{\hat{\mathbf{B}}}_{OPT}.\end{aligned}$$

and

$$e_{2k} = y_k - \mathbf{x}_k' \hat{\hat{\mathbf{B}}}_{OPT}.$$

4.2 Non-nested Double Sampling

We obtain the estimated variance of \hat{Y}_{OPT} by using the following approximation.

$$\begin{aligned}\tilde{Y}_{OPT} &= \hat{Y}_{HT} + (\mathbf{X}_1 - \tilde{\mathbf{X}}_1)' \tilde{\mathbf{B}}_{1,OPT} + (\tilde{\mathbf{X}} - \hat{\hat{\mathbf{X}}})' \tilde{\tilde{\mathbf{B}}}_{OPT} \\ &= \ddot{Y}_{OPT} + O_p(n_1^{-1/2}) \quad (4.8)\end{aligned}$$

where

$$\ddot{Y}_{OPT} = \hat{Y}_{HT} + (\mathbf{X}_1 - \tilde{\mathbf{X}}_1)' \mathbf{B}_{1,OPT} + (\tilde{\mathbf{X}} - \hat{\hat{\mathbf{X}}})' \mathbf{B}_{OPT}. \quad (4.9)$$

Decomposing \ddot{Y}_{OPT} into more elementary components, we have that:

$$\begin{aligned}\ddot{Y}_{OPT} &= \hat{Y}_{HT} + \left(\mathbf{X}_1 - \frac{\hat{\mathbf{X}}_1 + \alpha \hat{\hat{\mathbf{X}}}_1}{1 + \alpha} \right)' \mathbf{B}_{1,OPT} \\ &+ \frac{(\tilde{\mathbf{X}} - \hat{\hat{\mathbf{X}}})'}{1 + \alpha} \mathbf{B}_{OPT} \\ &= \left(\hat{Y}_{HT} - \frac{1}{1 + \alpha} (\alpha \hat{\hat{\mathbf{X}}}_1' \mathbf{B}_{1,OPT} + \hat{\hat{\mathbf{X}}}' \mathbf{B}_{1,OPT}) \right) \\ &+ \left(\mathbf{X}_1' \mathbf{B}_{1,OPT} - \frac{1}{1 + \alpha} (\hat{\mathbf{X}}_1' \mathbf{B}_{1,OPT} - \hat{\hat{\mathbf{X}}}' \mathbf{B}_{OPT}) \right). \quad (4.10)\end{aligned}$$

The variance of \ddot{Y}_{OPT} is:

$$\begin{aligned}V(\ddot{Y}_{OPT}) &= V \left(\hat{Y}_{HT} - \frac{1}{1 + \alpha} (\alpha \hat{\hat{\mathbf{X}}}_1' \mathbf{B}_{1,OPT} + \hat{\hat{\mathbf{X}}}' \mathbf{B}_{OPT}) \right) \\ &+ \frac{1}{(1 + \alpha)^2} [\alpha \mathbf{B}_{1,OPT}' V(\hat{\hat{\mathbf{X}}}_1) \mathbf{B}_{1,OPT} \\ &+ \mathbf{B}_{OPT}' V(\hat{\hat{\mathbf{X}}}) \mathbf{B}_{OPT} \\ &+ 2\alpha (\mathbf{B}_{OPT}' V(\hat{\hat{\mathbf{X}}}) \tilde{\mathbf{B}}_{1,OPT}' + \text{Cov}(\hat{\hat{\mathbf{X}}}_1, \hat{\hat{\mathbf{X}}}') \mathbf{B}_{OPT})] \quad (4.11)\end{aligned}$$

Result 3: The estimated variance of \tilde{Y}_{OPT} , $\hat{V}(\tilde{Y}_{OPT})$, defined by equation (4.8) is approximately equal to:

$$\begin{aligned}\hat{V}(\tilde{Y}_{OPT}) &= \hat{V} \left(\hat{Y}_{HT} - \frac{1}{1 + \alpha} (\alpha \hat{\hat{\mathbf{X}}}_1' \tilde{\mathbf{B}}_{1,OPT} + \hat{\hat{\mathbf{X}}}' \tilde{\tilde{\mathbf{B}}}_{OPT}) \right) \\ &+ \frac{1}{(1 + \alpha)^2} [\alpha \tilde{\mathbf{B}}_{1,OPT}' \hat{V}(\hat{\hat{\mathbf{X}}}_1) \tilde{\mathbf{B}}_{1,OPT} + \tilde{\tilde{\mathbf{B}}}_{OPT}' \hat{V}(\hat{\hat{\mathbf{X}}}) \tilde{\tilde{\mathbf{B}}}_{OPT} \\ &+ 2\alpha (\tilde{\tilde{\mathbf{B}}}_{OPT}' \hat{V}(\hat{\hat{\mathbf{X}}}) \tilde{\mathbf{B}}_{1,OPT} + \text{Cov}(\hat{\hat{\mathbf{X}}}_1, \hat{\hat{\mathbf{X}}}') \tilde{\tilde{\mathbf{B}}}_{OPT})]. \quad (4.12)\end{aligned}$$

Computation of the first term of (4.12) is based on the residuals $y_k - (\alpha \mathbf{x}_{1k}' \tilde{\mathbf{B}}_{1,OPT} + \mathbf{x}_k' \tilde{\tilde{\mathbf{B}}}_{OPT}) / (1 + \alpha)$. The computation of the other terms of (4.12) is mainly based on the estimated variances of $\hat{\hat{\mathbf{X}}}_1$ and of $\hat{\hat{\mathbf{X}}}$, as well as on their estimated covariances. We can use the approximation of the variance, as described by Tillé (2001), and suitably adapt it to estimate the required covariances.

5. SOME SPECIFIC EXAMPLES

Three traditional examples for double sampling are presented for the two cases (nested and non-nested). Furthermore, we briefly describe how two major business surveys carried out by Statistics Canada use double sampling.

5.1 Nested Sampling

Example 1: Let us assume that a simple random sample s_1 of size n_1 is selected from a population U of size N . The sample is stratified into L strata s_{1h} each of size n_{1h} . Random samples s_{2h} of size n_{2h} are then selected without replacement in each stratum s_{1h} . The estimator of the total is $\hat{Y}_{\text{EXP}} = N \sum_{h=1}^L p_{1h} \bar{y}_{2h} = N \bar{y}_{2, \text{st}}$, where $p_{1h} = n_{1h}/n_1$. Using (4.7), we can show that the estimated variance of \hat{Y}_{EXP} , $\hat{V}(\hat{Y}_{\text{EXP}})$, consists of the sum of $\hat{V}_1(\hat{Y}_{\text{EXP}})$ and $\hat{V}_2(\hat{Y}_{\text{EXP}})$ corresponding to the first and second phases of the sampling design. Thus:

$$\hat{V}(\hat{Y}_{\text{EXP}}) = \hat{V}_1(\hat{Y}_{\text{EXP}}) + \hat{V}_2(\hat{Y}_{\text{EXP}})$$

where

$$\hat{V}_1(\hat{Y}_{\text{EXP}}) = N^2 \frac{(1-f_1)}{n_1} \sum_{h=1}^L p_{1h} \left[(1-a_h) \hat{S}_{2yh}^2 + \frac{n_1}{n_1-1} (\bar{y}_{2h} - \bar{y}_{2, \text{st}})^2 \right];$$

$$\hat{V}_2(\hat{Y}_{\text{EXP}}) = N^2 \sum_{h=1}^L \frac{(1-f_{2h})}{n_{2h}} p_{1h}^2 \hat{S}_{2yh}^2;$$

and

$$a_h = \frac{(n_1 - n_{1h})}{n_{2h}(n_1 - 1)}; f_1 = \frac{n_1}{N}; f_{2h} = \frac{n_{2h}}{n_{1h}};$$

$$\hat{S}_{2yh}^2 = \frac{1}{n_{2h} - 1} \sum_{k \in s_{2h}} (y_k - \bar{y}_{2h})^2;$$

$$\bar{y}_{2h} = \frac{1}{n_{2h}} \sum_{k \in s_{2h}} y_k$$

$$\text{and } \bar{y}_{2, \text{st}} = \sum_{h=1}^L p_{1h} \bar{y}_{2h}.$$

Example 2: Let us assume that, for the sampling design described in Example 1, we also have auxiliary data, \mathbf{x}_k , available in the first phase s_1 . If we assume that the slopes (β_h) vary among the strata, we can assume that the following model $y_k = \mathbf{x}_k' \beta_h + \epsilon_k$ holds, where $E(\epsilon_k) = 0$, $E(\epsilon_k^2) = \sigma_k^2$, $k \in s_{1h}$, $h = 1, \dots, L$, and $E(\epsilon_k \epsilon_\ell) = 0$ for $k \neq \ell$, for $k, \ell \in s_{1h}$, $h = 1, \dots, L$. This model gives us a separate regression estimator, that is,

$$\hat{Y}_{\text{SEP, REG}} = \sum_{h=1}^L \frac{N}{n_1} \frac{n_{1h}}{n_{2h}} \sum_{k \in s_{2h}} g_{2k} y_k$$

where

$$g_{2k} = 1 + \left(\sum_{k \in s_{1h}} \mathbf{x}_k' - \sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} \mathbf{x}_k' \right) \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \frac{\mathbf{x}_k}{\sigma_k^2}$$

if $k \in s_{2h}$. In each stratum h , the slopes β_h are estimated as

$$\hat{B}_{2h} = \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \right).$$

The variance of $\hat{Y}_{\text{SEP, REG}}$ is estimated as being the sum of the variance components of each phase. These components are $\hat{V}_1(\hat{Y}_{\text{EXP}})$ and $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$, where $\hat{V}_1(\hat{Y}_{\text{EXP}})$ was defined in example 1. Variance $\hat{V}_2(\hat{Y}_{\text{SEP, REG}})$ is obtained by replacing variable y_k by $e_k = g_k(y_k - \mathbf{x}_k' \hat{B}_h)$ in $\hat{V}_2(\hat{Y}_{\text{EXP}})$. The estimated variance of $\hat{Y}_{\text{SEP, REG}}$ is therefore:

$$\hat{V}(\hat{Y}_{\text{SEP, REG}}) = \frac{N^2(1-f_1)}{n_1} \sum_{h=1}^L p_{1h} \left[(1-a_h) \hat{S}_{2yh}^2 + \frac{n_1}{n_1-1} (\bar{y}_{2h} - \bar{y}_{2, \text{st}})^2 \right] + \sum_{h=1}^L \frac{N^2(1-f_{2h})}{n_{2h}} p_{1h}^2 \hat{S}_{2eh}^2$$

where

$$\hat{S}_{2eh}^2 = \sum_{k \in s_{2h}} \frac{(e_k - \bar{e}_h)^2}{n_{2h} - 1}$$

and

$$\hat{S}_{2yh}^2 = \frac{1}{n_{2h} - 1} \sum_{k \in s_{2h}} (y_k - \bar{y}_{2h})^2.$$

5.2 Non-nested Sampling

These two examples are taken from Des Raj (1968, pages 142–149). We are using them to illustrate the results of sections 3 and 4. We consider two different sampling designs.

With the first sampling design, we assume that: (i) the first sample s_1 of size n_1 is selected with a simple random sampling design without replacement from population U ; and (ii) the second sample s_2 of size n_2 is selected either by using measurements of size x_i found in the first sample s_1 (nested case) or by selecting it independently (non-nested case) from the first sample s_1 in a manner proportional to

size x_i (known for all units of the population). The resulting estimator is

$$\hat{Y}_{\text{EPTAR}} = \frac{N}{n_1} \frac{\sum_{s_1} x_i}{n_2} \sum_{s_2} \frac{y_i}{x_i}.$$

For the second sampling design, we assume that the two samples s_1 and s_2 have been selected using a simple random sampling design without replacement. Here again, we examine the nested and non-nested cases. We assume that we find the auxiliary observation x_i for any unit selected in the first sample s_1 . The estimator is $\hat{Y}_{\text{RAT}} = (N/n_1 \sum_{s_1} x_i) (\sum_{s_2} y_i / \sum_{s_2} x_i) = \hat{X} \hat{R}$. Table 3 summarizes these two sampling designs, as well as this corresponding estimators with their estimated variances for the nested and non-nested cases.

The undefined terms in Table 3 are given by $p_{1i} = x_i / \sum_{s_1} x_i$; $p_i = x_i / \sum_U x_i$; $V(\hat{Y}_p) = 1/n_1 \sum_U p_i (y_i/p_i - Y)^2$; $S_{y-Rx} = (N-1)^{-1} \sum_U (y_i - R x_i)^2$; $f_2 = n_2/N$ $f_1 = n_1/N$, and $R = Y/X$.

Table 3 shows that there is little difference in the variances between the nested and non-nested cases. For \hat{Y}_{EPTAR} , the variance will be smaller for the nested case if the coefficient of variation (CV) of variable y is smaller than that of variable x . For \hat{Y}_{RAT} , the variance will be smaller for the nested case if $\rho \text{CV}(\bar{y}) < \text{CV}(\bar{x})$ where ρ is the correlation between y and x .

5.3 Two Statistics Canada Surveys

Several Statistics Canada surveys use double sampling. We will illustrate the ideas presented in this paper using two business surveys. These surveys are the Quarterly Retail Commodity Survey (QRCS) and the Survey of Employment, Payrolls and Hours (SEPH). The Quarterly Retail Commodity Survey uses nested double sampling, whereas the Survey of Employment, Payrolls and Hours (SEPH) uses non-nested double sampling.

The Quarterly Retail Commodity Survey: The purpose of the (QRCS) is to obtain detailed information on retail commodity sales on a quarterly basis. The RCS is a sub-sample of the Monthly Survey of Retail Trade (MRTS), a monthly survey. The MRTS measures mainly sales by trade group (group of three or four-digit codes of the 1980 Standard Industrial Classification (SIC)), by province and for certain census metropolitan areas (CMA). The target population is statistical companies with statistical locations identified on the Business Register and which are active in the retail trade. About 16,000 companies are interviewed each month. The population is stratified by province, territory, certain CMA and by trade group.

The MRTS is stratified in H strata, based on size (2-3 groups), geography (10 provinces, 2 territories) and industry (16 main groups). This sample is re-stratified independently for the QRCS. The QRCS stratification differs from the MRTS geographically, by size and by industry. A sub-sample is selected using the "new" stratification of the MRTS sample. The QRCS estimate is based on a double-ratio estimator that uses auxiliary data (sales) from the MRTS. The second-phase sampling unit (QRCS) remains the statistical company. The first-phase sample is re-stratified by trade group, by province and by size based on the most recent information from the MRTS. For stratification purposes, each company is assigned a province and a dominant trade group based on the one that generates the most sales. The two-phase estimator is used by the MRTS. Binder, Babyak, Brodeur, Hidirolglou, and Jocelyn (2000) derived a variance estimator that took into account the sampling design and the estimation method. They expressed variance estimators of the total as simple sums of appropriate residual terms for the case of the ratio estimator.

The results of Binder *et al.* (2000) can be adapted to incorporate the optimal regression estimator in each phase. We assume that the auxiliary information (x_{1k}) is known at

Table 3
Two Sampling Designs with Nested and Non-nested Samples

	Sampling design 1	Sampling design 2
Sampling Design	$N \rightarrow n_1$ (SRSWOR) $n_1 \rightarrow n_2$ (PPSWOR)	$N \rightarrow n_1$ (SRSWOR) $n_1 \rightarrow n_2$ (SRSWOR)
Estimator	$\hat{Y}_{\text{EPTAR}} = \frac{N}{n_1} \sum_{s_1} \frac{y_i}{n_2 p_{1i}}$	$\hat{Y}_{\text{RAT}} = \sum_{s_1} \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} = \hat{X} \hat{R}$
Variance		
Nested	$N^2 \frac{(1-f_1)}{n_1} S_y^2 + \frac{V(\hat{Y}_p)}{n_2}$	$\frac{N^2(1-f_1)}{n_1} (2R S_{xy} - R^2 S_x^2) + N^2 \frac{(1-f_2)}{n_2} S_{y-Rx}^2$
Non-nested	$N^2 \frac{(1-f_1)}{n_1} R^2 S_x^2 + \frac{V(\hat{Y}_p)}{n_2} \left[1 + \frac{1}{n_1} (1-f_1) \frac{S_x^2}{\bar{X}^2} \right]$	$\frac{N^2(1-f_1)}{n_1} R^2 S_x^2 + N^2 \frac{(1-f_2)}{n_2} S_{y-Rx}^2$

the level of population U , either for each unit $k \in U$ or for the total $X_{1k} = \sum_U x_{1k}$. The QRCS sampling design can be formally stated as follows. The population is stratified in H strata U_h ; $h = 1, \dots, H$, and simple random samples without replacement s_{1h} , of size n_{1h} , are selected in each stratum U_h . The x_k variable is observed for each unit belonging to s_1 . The resulting first-phase sample, $s_1 = U_{h=1}^H s_{1h}$, is then stratified in strata s_{1g} , $g = 1, \dots, G$. The stratification of s_1 is independent of the stratification of the universe U . A simple random sample s_{2g} of size n_{2g} is then selected from each stratum s_{1g} , $g = 1, \dots, G$. We observe (y_k, x'_k) , where $x_k = (x'_{1k}, x'_{2k})'$ for each unit belonging to sample $s_2 = U_{g=1}^G s_{2g}$. We assume that models $y_k = x'_{1k} \beta_1 + \varepsilon_{1k}$ and $y_k = x'_k \beta + \varepsilon_{2k}$ hold for s_1 and s_2 respectively. For each of these models $\varepsilon_{1k} \sim (0, \sigma^2_{z_{1k}})$ and $\varepsilon_{2k} \sim (0, \sigma^2_{z_{2k}})$ where z_{1k} and z_{2k} are known positive factors. If $z_{1k} \neq 1$ or $z_{2k} \neq 1$ for all units $k \in U$, the data can be standardized by dividing them either by $\sqrt{z_{1k}}$ or $\sqrt{z_{2k}}$. The resulting optimal regression estimator for the total Y is given by:

$$\tilde{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (X_1 - \tilde{X}_1)' \tilde{B}_{1, \text{OPT}} + (\hat{X} - \tilde{X})' \tilde{B}_{\text{OPT}}$$

where the components of \tilde{Y}_{OPT} were defined in section 3.1. The simplified form (without double sums) of the variance of \tilde{Y}_{OPT} is:

$$\begin{aligned} \hat{V}(\tilde{Y}_{\text{OPT}}) = & \sum_{h=1}^H N_h^2 (1 - f_{1h}) \frac{\hat{S}_{1h}^2}{n_{1h}} \\ & + \sum_{g=1}^G n_{1g}^2 (1 - f_{2g}) \frac{\hat{S}_{2g}^2}{n_{2g}} \\ & + \sum_{h=1}^H \sum_{g=1}^G \frac{N_h^2 (1 - f_{1h}) n_{2g}^2 (1 - f_{2g})}{n_{1h}^2 (n_{1h} - 1)} \frac{\hat{S}_{2hg}^2}{n_{2h}} \end{aligned}$$

where the variances are defined by

$$\begin{aligned} \hat{S}_{1h}^2 = & \frac{1}{n_{1h} - 1} \left\{ \sum_{g=1}^G \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k}^2 - \frac{1}{n_{1h}} \left(\sum_{g=1}^G \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k} \right)^2 \right\}; \\ \hat{S}_{2hg}^2 = & \frac{1}{n_{2hg} - 1} \sum_{k=1}^{n_{2hg}} (\tilde{e}_{1k} - \bar{\tilde{e}}_{1(hg)})^2 \end{aligned}$$

and

$$\hat{S}_{2g}^2 = \frac{1}{n_{2g} - 1} \sum_{k=1}^{n_{2g}} (\tilde{e}_{2k} - \bar{\tilde{e}}_{2h})^2.$$

The means in these estimated variances are

$$\bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}, \quad \bar{\tilde{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}$$

and

$$\bar{\tilde{e}}_{2h} = \frac{1}{n_{2g}} \sum_{k=1}^{n_{2g}} \tilde{e}_{2k}.$$

Here, n_{2hg} is the number of units selected in sample s_2 belonging to the intersection of strata U_h and s_{1g} . Also, the required residuals are $\tilde{e}_{1k} = g_{1k}(y_k - x'_{1k} \tilde{B}_{1, \text{OPT}})$ and $\tilde{e}_{2k} = g_{2k}(y_k - x'_k \tilde{B}_{\text{OPT}})$. The adjustment factors g_{1k} and g_{2k} are as defined in section 4.1.

The Survey of Employment, Payrolls and Hours: The objective of this survey is to obtain estimates of the number of paid employees, the average weekly payroll and other related variables using various combinations of industry and province. This survey was recently redesigned to use administrative data for all businesses included in the survey universe. The survey produces estimates based on both the administrative data (ADMIN sample) and data directly obtained by a survey known as the Business Payroll Survey (BPS).

The ADMIN sample s_1 consists of some 200,000 units selected from universe U_1 of the pay deduction accounts to obtain the administrative data. The sampling design for this sample is stratified Bernoulli (by region), and the sampling rate varies between 10% to 100% amongst the different strata (region). The size of the sample represents approximately 20% of the total number of pay deduction accounts. Only two variables represented as $(x_{1k}^{(1)})$ are available from the administrative source: these are the number of paid employees and the gross monthly payroll.

The BPS sample s_2 consists of approximately 10,000 establishments drawn from the Business Register U_2 . The BPS collects the same two variables as the administrative source, namely, the number of paid employees and the gross monthly payroll denoted as $(x_{1k}^{(2)})$, several other variables $(x_{2k}^{(2)})$ of interest defined by type of employee (employees paid by the hour, salaried, active owners, other employees), and variables of interests, such as the number of paid hours and weekly earnings, $(y_k^{(2)})$. More information on the BPS is provided in Rancourt and Hidiogrou (1998).

The BPS is stratified by industry type, geographic region and size (varying from two to three groups based on the number of employees). These strata were designed to take into account the different regression models between $y_k^{(2)}$ and $x_k^{(2)}$. The resulting estimated regression coefficients are used to predict \hat{y}_k for each sampled administrative record. There are two steps involved in the estimation of the total for a given variable of interest. First, the sampling weights $w_k^{(1)}$ associated with the administrative data are calibrated using known regional population counts, N_i , for regions U_{1i} , $i = 1, \dots, I$. The adjusted weight of a sample unit k belonging to region U_{1i} is $\bar{w}_k^{(1)} = w_k^{(1)} g_{1i}$, where $g_{1i} = N_i / \sum_{s_{1i}} w_k^{(1)}$ and $s_{1i} = s_1 \cap U_{1i}$. Second, $y_k^{(2)}$ is regressed on $x_k^{(2)}$ using subsets $s_{2,j}$, $j = 1, \dots, J$, of the s_2 sample. The $s_{2,j}$

subsets, classified by industry, region and sometimes size, are formed in advance to obtain the best possible regression fits. For each subset $s_{2,j}$, the estimated regression vectors \hat{B}_j are obtained as:

$$\hat{B}_j = \left(\sum_{s_{2,j}} w_k^{(2)} x_k^{(2)} x_k'^{(2)} / \hat{\sigma}_k^2 \right)^{-1} \sum_{s_{2,j}} w_k^{(2)} x_k^{(2)} y_k^{(2)} / \hat{\sigma}_k^2;$$

$$j = 1, \dots, J$$

where $w_k^{(2)}$ is the sampling weight for each sampled establishment, and $\hat{\sigma}_k^2$ are known positive factors that control the impact of outliers or define the required estimator. For example, if $\hat{\sigma}_k^2$ is proportional to one of the components of $x_k^{(2)}$, we obtain the ratio estimator. The estimator of total for a variable y is therefore $\hat{Y} = \sum_{j=1}^J \sum_{s_{1,j}} \tilde{w}_k^{(1)} x_k'^{(1)} \hat{B}_j$, where $s_{1,j}$ is a partition of s_1 corresponding to the subsets defining $s_{2,j}$. SEPH is an example of a non-nested double sampling sampling design. More details of the SEPH redesign are available in Hidirolou (1995) and Hidirolou, Latouche, Armstrong and Gossen (1995).

6. CONCLUSION

Nested and non-nested double sampling are usually treated separately in the literature. Given that the population total Y is of interest, and that there is auxiliary information available, this paper has unified the estimation procedures for these two sampling methods using an optimal regression approach. Also, for the nested case, the procedure has been linked to the GREG procedure proposed by Hidirolou and Särndal (1998). For the non-nested case, the method used by Deville (1999) has been extended when there are also auxiliary data at the population level. Lastly, practical examples were provided to illustrate this theory.

REFERENCES

- BERGER, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.
- BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A. and JOCELYN, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 4, 751-764.
- BREIDT, J., and FULLER, W.A. (1993). Regression weighting for multiphase samples. *Sankhya*, 55, 297-309.
- BREWER, K. (2000). Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion probabilities. In the *Proceedings of the Second International Conferences on Establishment Surveys*. Buffalo, New York, 1417-1422.
- CASSADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimation under normal theory. *Survey Methodology*, 19, 183-192.
- CHAUDHURI, A., and ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley and Sons.
- DES RAJ (1968). *Sampling Theory*. TMH Edition.
- DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology*, 25, 193-204.
- DEVILLE, J.-C. (1999). Simultaneous calibrating of several surveys. *Proceedings: Symposium 1999, Combining Data from Different Sources*, 207-212.
- HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the canadian survey of employment, payrolls and hours survey redesign. *Proceedings of The Survey Methods Section*, Statistical Society of Canada, 123-128.
- HIDIROGLOU, M.A., LATOUCHE, M., ARMSTRONG, B. and GOSSEN, M. (1995). Improving survey information using administrative records: The case of the canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.
- HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- KORN, E.L., and GRAUBARD, B.I. (1999). *Analysis of Health Surveys*. Wiley series in probability and Statistics.
- MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.
- MONTANARI, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69-77.
- MONTANARI, G.E. (2000). Conditioning on auxiliary variables means in finite population inference. *Australian New Zealand Journal of Statistics*, 42, 407-421.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- RANCOURT, E., and HIDIROGLOU, M.A. (1998). Use of administrative records in the Canadian survey of employment, payrolls and hours. *Proceedings of the Survey Methods Section*, 39-47.
- RAO, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 60, 125-133.
- RAO, J.N.K. (1994). Estimation of totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-166.
- RÖSEN, B. (2000). A user's guide to pareto π ps sampling. In the *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, New York, 289-294.
- SÄRNDAL, C.E. (1996). Efficient estimators with simple variances in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, Y. (1992). *Model assisted survey sampling*. New York, Springer-Verlag.
- TAM, S. M. (1984). On covariances from nested samples. *The American Statistician*, 38, 288-289.
- TILLÉ, Y. (2001) *Théorie des Sondages : Échantillonnage et estimation en population finies*. Dumond.

Estimation Using the Generalised Weight Share Method: The Case of Record Linkage

PIERRE LAVALLÉE and PIERRE CARON¹

ABSTRACT

More and more, databases are combined using record linkage methods to increase the amount of available information. When there is no unique identifier to perform the matching, a probabilistic linkage is used. A record on the first file is linked to a record on the second file with a certain probability, and then a decision is made on whether this link is a true link or not. This process usually requires a certain amount of manual resolution that is costly in terms of time and employees. Also, this process often leads to a complex linkage. That is, the linkage between the two databases is not necessarily one-to-one, but can rather be many-to-one, one-to-many, or many-to-many.

Two databases combined using record linkage can be seen as two populations linked together. We consider in this paper the problem of producing estimates for one of the populations (the target population) using a sample selected from the other one. We assume that the two populations have been linked together using probabilistic record linkage. To solve the estimation problem issued from a complex linkage between the population where the sample is selected and the target population, Lavallée (1995) suggested the use of the Generalised Weight Share Method (GWSM). This method is an extension of the Weight Share Method presented by Ernst (1989) in the context of longitudinal household surveys.

The paper will first provide a brief overview of record linkage. Secondly, the GWSM will be described. Thirdly, the GWSM will be adapted to provide three different approaches that take into account linkage weights issued from record linkage. These approaches will be: (1) use all non-zero links with their respective linkage weights; (2) use all non-zero links above a given threshold; and (3) choose the links randomly using Bernoulli trials. For each of the approaches, an unbiased estimator of a total will be presented together with a variance formula. Finally, some simulation results that compare the three proposed approaches to the Classical Approach (where the GWSM is used based on links established through a decision rule) will be presented.

KEY WORDS: Generalised weight share method; Record linkage; Estimation; Clusters.

1. INTRODUCTION

To augment the amount of available information, data from different sources are increasingly being combined. These databases are often combined using record linkage methods. When the files involved have a unique identifier that can be used, the linkage is done directly using the identifier as a matching key. When there is no unique identifier, a probabilistic linkage is used. In that case, a record on the first file is linked to a record on the second file with a certain probability, and then a decision is made on whether this link is a true link or not. Note that this process usually requires a certain amount of manual resolution that is costly in terms of time and employees.

We consider the production of an estimate of a total (or a mean) of one target clustered population when using a sample selected from another population linked to the first population. We assume that the two populations have been linked together using probabilistic record linkage. Note that this type of linkage often leads to a complex linkage between the two populations. That is, the linkage between the units of each of the two populations is not necessarily one-to-one, but can rather be many-to-one, one-to-many, or many-to-many.

To solve the estimation problem caused by a complex linkage between the population where the sample is selected and the target population, Lavallée (1995) suggested the use of the Generalised Weight Share Method (GWSM). This method is an extension of the Weight Share Method presented by Ernst (1989). Although this last method has been developed in the context of longitudinal household surveys, it was shown that the Weight Share Method can be generalised to situations where a target population of clusters is sampled through the use of a frame which refers to a different population, but somehow linked to the first one.

The problem that is considered in this paper is to estimate the total of a characteristic of a target population that is naturally divided into clusters. Assuming that the sample is obtained by the selection of units within clusters, if at least one unit of a cluster is selected, then the whole cluster is interviewed. This usually leads to cost reductions as well as the possibility of producing estimates on the characteristics of both the clusters and the units.

In the present paper, we will try to answer the following questions:

- a) Can we use the GWSM to handle the estimation problem related to populations linked together through record linkage?

¹ Pierre Lavallée and Pierre Caron, Statistics Canada, Business Survey Methods Division, Ottawa, Ontario, K1A 0T6, e-mail: plavall@statcan.ca and caropie@statcan.ca.

- b) Can we adapt the GWSM to take into account the linkage weights issued from record linkage?
- c) Can GWSM help in reducing the manual resolution required by record linkage?
- d) If there is more than one approach to use the GWSM, is there a “better” approach?

It will be seen that the answer is clearly yes to (a) and (b). However, for question (c), it will be shown that there is a price to pay in terms of an increase to the sample size, and therefore to the collection costs. For question (d), although there is no definite answer, some approaches seem to generally be more appropriate.

The paper will first provide a brief overview of record linkage. Secondly, the GWSM will be described. Thirdly, the GWSM will be adapted to provide three different approaches that take into account linkage weights issued from record linkage. These approaches will be: (1) use all non-zero links with their respective linkage weights; (2) use all non-zero links above a given threshold; and (3) choose the links randomly using Bernoulli trials. For each of the approaches, an unbiased estimator of a total will be presented together with a variance formula. Finally, some simulation results that compare the three proposed approaches to the Classical Approach (where the GWSM is used based on links established through a decision rule) will be presented.

2. RECORD LINKAGE

The concepts of record linkage were introduced by Newcome, Kennedy, Axford and James (1959) and formalised in the mathematical model of Fellegi and Sunter (1969). As described by Bartlett, Krewski, Wang and Zielinski (1993), record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same unit (individual or business). Record linkage is sometimes also called exact matching, in contrast to statistical matching. This last process attempts to link files that have few units in common (see Budd and Radner 1969, Budd 1971, Okner 1972, and Singh, Mantel, Kinack and Rowe 1993). With statistical matching, linkages are based on similar characteristics rather than unique identifying information. In the present paper, we will restrict ourselves to the context of record linkage. However, the developed theory could also be used for statistical matching.

Suppose that we have two files A and B containing characteristics relating to two populations U^A and U^B , respectively. The two populations are somehow related to each other. They can represent, for example, exactly the same population, where each of the files contains a different set of characteristics of the units of that population. They can also represent different populations, but with some natural links between them. For example, one population

can be one of parents, and the other population one of children belonging to the parents. Note that the children usually live in households that can be viewed as clusters. Another example is one of an agricultural survey where the first population is a list of farms as determined by the Canadian Census of Agriculture and the second population is a list of taxation records from the Canadian Customs and Revenue Agency (CCRA). In the first population, each farm is identified by a unique identifier called the FarmID and some additional variables such as the name and address of the operators that are collected through the Census questionnaire. The second population consists of taxation records of individuals who have declared some form of agricultural income. These individuals live in households. The unique identifier on those records is either a social insurance number or a corporation number depending on whether or not the business is incorporated. However, each income tax report submitted to CCRA contains similar variables (name and address of respondent, *etc.*) as those collected by the Census.

The purpose of record linkage is to link the records of the two files A and B. If the records contain unique identifiers, then the matching process is trivial. For example, in the agriculture example, if both files would contain the FarmID, the matching process could be done using a simple matching procedure. Unfortunately, often a unique identifier is not available and then the linkage process needs to use some probabilistic approach to decide whether two records of the two files are linked together or not. With this linkage process, the likelihood of a correct match is computed and, based on the magnitude of this likelihood, it is decided whether we have a link or not.

Formally, we consider the product space $A \times B$ from the two files A and B. Let j indicate a record (or unit) from file A (or population U^A) and k a record (or unit) from file B (or population U^B). For each pair (j, k) of $A \times B$, we compute a linkage weight reflecting the degree to which the pair (j, k) is likely to be a true link. The higher the linkage weight is, the more likely the pair (j, k) is a true link. The linkage weight is commonly based on the ratios of the conditional probabilities of having a match μ and an unmatched $\bar{\mu}$ given the result of the outcome of the comparison C_{qjk} of the characteristic q of the records j from A and k from B, $q = 1, \dots, Q$. That is,

$$\begin{aligned} \hat{\theta}_{jk} &= \log_2 \left\{ \frac{P(\mu_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})}{P(\bar{\mu}_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})} \right\} \\ &= \hat{\theta}_{1jk} + \hat{\theta}_{2jk} + \dots + \hat{\theta}_{Qjk} + \hat{\theta}_{*jk} \end{aligned} \quad (2.1)$$

where $\hat{\theta}_{qjk} = \log_2 \left\{ \frac{P(C_{qjk} | \mu_{jk})}{P(C_{qjk} | \bar{\mu}_{jk})} \right\}$ for $q = 1, \dots, Q$, and

$$\hat{\theta}_{*jk} = \log_2 \left\{ \frac{P(\mu_{jk})}{P(\bar{\mu}_{jk})} \right\}.$$

The mathematical model proposed by Fellegi and Sunter (1969) takes into account the probabilities of an error in the linkage of units j from A and k from B. The linkage weight is then defined as

$$\theta_{jk}^{FS} = \sum_{q=1}^Q \theta_{qjk}^{FS}$$

where

$$\theta_{qjk}^{FS} = \begin{cases} \log_2 & \text{if characteristic } q \text{ of pair } (j,k) \text{ agrees} \\ \log_2 ((1 - \eta_{qjk}) / (1 - \bar{\eta}_{qjk})) & \text{otherwise} \end{cases}$$

with $\eta_{qjk}^{FS} = P(\text{characteristic } q \text{ agrees} \mid \mu_{jk})$ and $\bar{\eta}_{qjk} = P(\text{characteristic } q \text{ agrees} \mid \bar{\mu}_{jk})$. Note that the definition of θ_{qjk}^{FS} assumes that the Q comparisons are independent.

The linkage weights given by (2.1) are defined on \mathbf{R} , the set of real numbers, i.e., $\theta_{jk} \in] -\infty, +\infty[$. When the ratio of the conditional probabilities of having a match μ and an unmatched $\bar{\mu}$ is equal to 1, we get $\theta_{jk} = 0$. When this ratio is close to 0, θ_{jk} tends to $-\infty$. It might then be more convenient to define the linkage weights on $[0, +\infty[$. This can be achieved by taking the antilogarithm of θ_{jk} . We then obtain the following linkage weight θ_{jk} :

$$\theta_{jk} = \frac{P(\mu_{jk} \mid C_{1jk} C_{2jk} \dots C_{Qjk})}{P(\bar{\mu}_{jk} \mid C_{1jk} C_{2jk} \dots C_{Qjk})}. \quad (2.2)$$

Note that the linkage weight θ_{jk} is equal to 0 when the conditional probabilities of having a match μ is equal to 0. In other words, we have $\theta_{jk} = 0$ when the probability of having a true link for (j, ik) is nul.

Once a linkage weight θ_{jk} has been computed for each pair (j, k) of $A \times B$, we need to decide whether the linkage weight is sufficiently large to consider the pair (j, k) a link. This is typically done using a decision rule. With the approach of Fellegi and Sunter, we use an upper threshold θ_{High} and a lower threshold θ_{Low} to which each linkage weight θ_{jk} is compared. The decision is made as follows:

$$D(j, k) = \begin{cases} \text{link} & \text{if } \theta_{jk} \geq \theta_{\text{High}} \\ \text{can be a link} & \text{if } \theta_{\text{Low}} < \theta_{jk} < \theta_{\text{High}} \\ \text{nonlink} & \text{if } \theta_{jk} \leq \theta_{\text{Low}}. \end{cases} \quad (2.3)$$

The lower and upper thresholds θ_{Low} and θ_{High} are determined by *a priori* error bounds based on false links and false nonlinks. When applying decision rule (2.3), some clerical decisions are needed for those linkage weights falling between the lower and upper thresholds. This is generally done by looking at the data, and also by using auxiliary information. In the agriculture example, variables such as date of birth, street address and postal code, which are available on both sources of data, can be used for this purpose. By being automated and also by working on a probabilistic basis, some errors can be introduced in the record linkage process. This has been discussed in several

papers, namely Bartlett *et al.* (1993), Belin (1993) and Winkler (1995).

The application of decision rule (2.3) leads to the definition of an indicator variable $l_{jk} = 1$ if the pair (j, k) is considered to be a link, and 0 otherwise. As for the decisions that need to be taken for those linkage weights falling between the lower and upper thresholds, some manual intervention may be needed to decide on the validity of the links. In the case where the files A and B represent the same population (with a different set of characteristics), it is likely that for each unit j from file A, there will be only one unit linked in file B. That is, the units should be linked on a one-to-one basis. Note that decision rule (2.3) does not prevent the existence of many-to-one, one-to-many, or many-to-many links. As mentioned before, because of the probabilistic aspect of the record linkage process, which might introduce some errors, there could be more than one link per unit. In practice, this problem is usually solved by some manual intervention. In the agriculture example, it can occur that multiple operators of a farm each submit a tax report to CCRA for the same farm (one-to-many). Similarly, an operator who runs more than one farm could submit only one income tax report for his operations (many-to-one). Finally, one can imagine a scenario of many-to-many links when an operator runs more than one farm, where each farm has a number of different operators. These situations can be represented by Figure 1. In Figure 1, unit $j=1$ of U^A has a one-to-one link to unit $k=1$ of U^B ; unit $j=2$ forms to a one-to-many link to units $k=2$ and $k=4$; and units $j=2$ and $j=3$ together form a many-to-one link to unit $k=4$. For the agriculture example, it is clear that deciding on the validity of the links is more difficult than the case of the same population since the former allows the possibility of having true many-to-one or one-to-many situations.

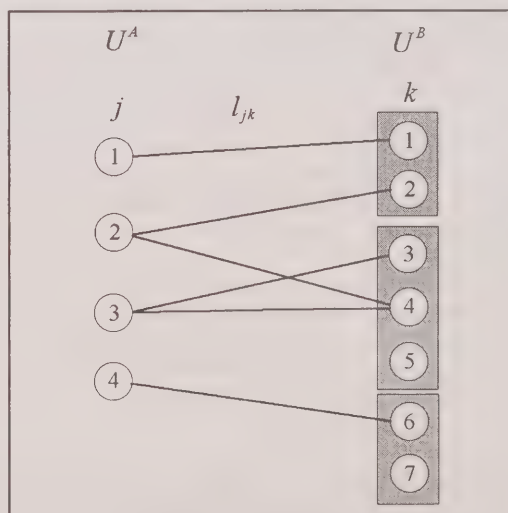


Figure 1. Example of links

3. THE GENERALISED WEIGHT SHARE METHOD

The GWSM is described in Lavallée (1995). It is an extension of the Weight Share Method described by Ernst (1989) but in the context of longitudinal household surveys. Various implications of using the Weight Share Method for longitudinal household surveys have been described by Gailly and Lavallée (1993). The GWSM can be viewed as a generalisation of *Network Sampling* and also of *Adaptive Cluster Sampling*. These two sampling methods are described in Thompson (1992), and Thompson and Seber (1996).

Suppose that a sample s^A of m^A units is selected from the population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit j . We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let the population U^B contain M^B units. This population is divided into N clusters where cluster i contains M_i^B units. For example, in the context of social surveys, the clusters can be households and the units can be the persons within the households. For business surveys, the clusters can be enterprises and the units can be the establishments within the enterprises. For the agriculture example, the clusters can be households, and the units, persons within the household who file an income tax report to CCRA.

We suppose that there exists a link between the units j of population U^A and the units k of clusters i of the population U^B . This link is identified by an indicator variable $l_{j,ik}$, where $l_{j,ik} = 1$ if there exists a link between unit $j \in U^A$ and unit $ik \in U^B$, and 0 otherwise. Note that there might be some units j of population U^A for which there is no link with any unit k of a cluster i of population U^B , i.e., $L_j^A = \sum_{i=1}^N \sum_{k=1}^{M_i^B} l_{j,ik} = 0$ for all $j \in U^A$. Also, there can be zero, one or more links for any unit k of a cluster i of population U^B , i.e., $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 0$, $L_{ik} = 1$ or $L_{ik} > 1$ for any $k \in U^B$.

With the GWSM, we have the following constraint:

Each cluster i of U^B must have at least one link with a unit j of U^A , i.e., $L_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$.

This constraint is essential for the GWSM to produce unbiased estimates. We will see in section 4 that in the context of record linkage, this constraint might not be satisfied.

For each unit j selected in s^A , we identify the units ik of U^B that have a non-zero link with j , i.e., $l_{j,ik} = 1$. For each identified unit ik , we suppose that we can establish the list of the M_i^B units of cluster i containing this unit. Then, each cluster i represents by itself a population U_i^B where $U^B = \bigcup_{i=1}^N U_i^B$. Let Ω^B be the set of the n clusters identified by the units $j \in s^A$.

From population U^B , we are interested in estimating the total $Y^B = \sum_{i=1}^N \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic y . An important constraint that is imposed in the measurement (or interviewing) process of y is to consider all units within the

same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit is interviewed. This constraint is one that often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. As an example, for social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example. For the agriculture example, one value of interest is the total farm revenue per household. In that case, we need to interview all persons within the household.

By using the GWSM, we want to assign an estimation weight w_{ik} to each unit k of an interviewed cluster i . To estimate the total Y^B belonging to population U^B , one can then use the estimator

$$\hat{Y} = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \quad (3.1)$$

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i . With the GWSM, the estimation process uses the sample s^A together with the links existing between U^A and U^B to estimate the total Y^B . The links are in fact used as a bridge to go from population U^A to population U^B , and vice versa.

The GWSM allocates to each interviewed unit ik a final weight established from an average of weights calculated within each cluster i entering into \hat{Y} . An *initial weight* that corresponds to the inverse of the selection probability is first obtained for all units k of cluster i of \hat{Y} having a non-zero link with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *final weight* is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster. This final weight is finally assigned to all units within the cluster. Note that the fact of allocating the same estimation weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into \hat{Y} is assigned an initial weight w'_{ik} as follows:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A} \quad (3.2)$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit ik having no link with any unit j of U^A has automatically an initial weight of zero. The final weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \quad (3.3)$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of U^B . The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ then corresponds to the total number of links present in cluster i . Finally, we assign $w_{ik} = w_i$ for all $k \in U_i^B$ and use equation (3.1) to estimate the total Y^B .

Using this last expression, it was shown in Lavallée (1995) that the GWSM is design unbiased. Further, let $z_{ik} = Y_i/L_i$ for all $k \in i$, where $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$. Then, \hat{Y} can be expressed as

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^N \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \quad (3.4)$$

and the variance of \hat{Y} is given by

$$\text{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \quad (3.5)$$

where $\pi_{jj'}^A$ is the joint probability of selecting units j and j' . See Särndal, Swensson and Wretman (1992) for the calculation of $\pi_{jj'}^A$ under various sampling designs. The variance $\text{Var}(\hat{Y})$ may be unbiasedly estimated from the following equation:

$$\text{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} t_j Z_j t_{j'} Z_{j'}. \quad (3.6)$$

Another unbiased estimator of the variance $\text{Var}(\hat{Y})$ may be developed in the form of Yates and Grundy (1953).

In presenting the Weight Share Method in the context of longitudinal surveys, Ernst (1989) proposed the use of constants α in the definition of the estimation weights. In the general context of the GWSM, the use of the same type of constants can be proposed. Let us define $\alpha_{j,ik} \geq 0$ for all pairs (j, ik) , with $\alpha_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \alpha_{j,ik} = 1$. We can then obtain new estimation weights as follows. For each unit k of cluster i entering into \hat{Y} , assign the following initial weight w_{ik}^{α} :

$$w_{ik}^{\alpha} = \sum_{j=1}^{M^A} \alpha_{j,ik} \frac{t_j}{\pi_j^A}. \quad (3.7)$$

The final weight w_i^{α} is given by

$$w_i^{\alpha} = \sum_{k=1}^{M_i^B} w_{ik}^{\alpha} = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} \alpha_{j,ik} \frac{t_j}{\pi_j^A}. \quad (3.8)$$

Finally, we assign $w_{ik}^{\alpha} = w_i^{\alpha}$ for all $k \in U_i^B$ and use equation (3.1) to estimate the total Y^B .

In the context of longitudinal surveys, Ernst (1989) noted that the most common choice for the constants α is the one where each individual receives one of two values: 0, or a non-zero value that is equal for all the remaining units within the cluster. In the present context, this would mean

to let $\alpha_{j,ik} = 0$ for all j and k in a subset U_i^{0B} of U_i^B , say, and $\alpha_{j,ik} = \text{constant}$ for all j and k in the complement subset $U_i^{\bar{0}B}$. Back to the context of longitudinal surveys, Kalton and Brick (1995) looked at the determination of optimal values for the α of Ernst (1989) where the optimality is measured in terms of minimal variance. They concluded that: "in the two-household case, the equal household weighting scheme minimises the variance of the household weights around the inverse selection probability weight when the initial sample is an equal epsem (equal probability) one." They also added that "in the case of an approximately epsem sample, the equal household weighting scheme should be close to the optimal, at least for the case where the members of the household at time t come from one or two households at the initial wave." This suggests that, for the GWSM, the choice of letting the constants α being 0 for some units and a positive value that is equal for all the remaining units within the cluster should be close to the optimal.

4. THE GWSM AND RECORD LINKAGE

With record linkage, the links $l_{j,ik}$ are established between files A and B, or population U_i^A and population U^B , using a probabilistic process. As mentioned before, record linkage uses a decision rule D such as (2.3) to decide whether there is a link or not between unit j from file A and unit ik from file B. Once the links are established, we then have the two populations U^A and U^B linked together, with the links identified by the indicator variable $l_{j,ik}$. Note that the decision rule (2.3) does not prevent the existence of complex links (many-to-one, one-to-many, or many-to-many).

Although the links can be complex, the GWSM can be used to estimate the total Y^B from population U^B using a sample s^A obtained from population U^A . Therefore, the answer is yes to question (a) stated in the introduction. Note that the estimates produced by the application of the GWSM might however not be unbiased if the constraint mentioned in section 3 is not satisfied. In that case, the use of the estimation weight (3.3) underestimates the total Y^B . To solve this problem, one practical solution is to collapse two clusters in order to get at least one non-zero link $l_{j,ik}$ for cluster i . This solution usually requires some manual intervention. Another solution is to impute a link by choosing one link at random within the cluster, or to choose the link with the largest linkage weight $\theta_{j,ik}$. Note that it might also happen that for a unit j of U^A , there is no non-zero link $l_{j,ik}$ with any unit ik of U^B . This is however not a problem since the only coverage in which we are interested is the one of U^B .

It is now clear that the GWSM can be used in the context of record linkage. The GWSM with the populations U^A and U^B linked together using record linkage with the decision rule (2.3) will be referred to as the Classical Approach.

Now, with the Classical Approach, the use of the GWSM is based on links identified by the indicator variable $l_{j,ik}$. Is it necessary to establish whether there is positively a link for each pair (j, ik) , or not? Would it be easier to simply use the linkage weights $\theta_{j,ik}$ (without using any decision rule) to estimate the total Y^B from U^B using a sample from U^A ? These questions lead to question (b) on whether or not it is possible to adapt the GWSM to take into account the linkage weights θ issued from record linkage.

In the present section, we will see that the answer to question (b) is yes by providing three approaches where the GWSM uses the linkage weights θ . The first approach is to use all the non-zero links identified through the record linkage process, together with their respective linkage weights θ . The second approach is the one where we use all the non-zero links with linkage weights above a given threshold θ_{High} . The third approach is one where the links are randomly chosen with probabilities proportional to the linkage weights θ .

4.1 Approach 1: Using all Non-Zero Links With Their Respective Linkage Weights

When using all non-zero links with the GWSM, one might want to give more importance to links that have large linkage weights θ , compared to those that have small linkage weights. By definition, for each pair (j, ik) of $A \times B$, the linkage weight $\theta_{j,ik}$ reflects the degree to which the pair (j, ik) is likely to be a true link. We then no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not between unit j from U^A and unit k of cluster i from U^B . Instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process. (This assumes that the file with the linkage weights is available. In practice, the only available file is often the linked file obtained at the end of the linkage process, once some manual resolution has been performed. In this case, the linkage weights are no longer available and the three proposed approaches to be used with the GWSM are immaterial to reduce the problem of manual resolution). Note that by doing so, we do not need any decision to be taken to establish whether there is a link or not between two units.

For each unit j selected in s^A , we identify the units ik of U^B that have a non-zero linkage weight with unit j , i.e., $\theta_{j,ik} > 0$. Let $\Omega^{\text{RL},B}$ be the set of the n^{RL} clusters identified by the units $j \in s^A$, where “RL” stands for “Record Linkage”. Note that because we use all non-zero linkage weights, we have $n^{\text{RL}} \geq n$. We now obtain the initial weight w_{ik}^{RL} by directly replacing the indicator variable l in equations (3.2) and (3.3) by the linkage weight θ .

$$w_{ik}^{\text{RL}} = \sum_{j=1}^{M^A} \theta_{j,ik} \frac{t_j}{\pi_j^A}. \quad (4.1)$$

The final weight w_i^{RL} is given by

$$w_i^{\text{RL}} = \frac{\sum_{k=1}^{M_i^B} w_{ik}^{\text{RL}}}{\sum_{k=1}^{M_i^B} \Theta_{ik}} \quad (4.2)$$

where $\Theta_{ik} = \sum_{j=1}^{M^A} \theta_{j,ik}$. Finally, we assign $w_{ik}^{\text{RL}} = w_i^{\text{RL}}$ for all $k \in U_i^B$. Note that by being present both at the numerator and denominator of equation (4.2), the linkage weights $\theta_{j,ik}$ do not need to be between 0 and 1. They just need to represent the relative likelihood of having a link between two units from populations U^A and U^B . It is also interesting to note that by letting $\alpha_{j,ik} = \theta_{j,ik} / \Theta_{ik}$, where $\Theta_{ik} = \sum_{j=1}^{M^A} \theta_{j,ik}$, we obtain, for the estimation weight w_i^{RL} , an equivalent formulation to the one given by (3.7) and (3.8).

With the Classical Approach, we stated the constraint that each cluster i of U^B must have at least one link with a unit j of U^A , i.e., $L_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$. This constraint is translated here into the need of having for each cluster i of U^B at least one non-zero linkage weight $\theta_{j,ik}$ with a unit j of U^A , i.e., $\Theta_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik} > 0$. In theory, the record linkage process does not insure that this constraint is satisfied. It might then turn out that for a cluster i of U^B , there is no non-zero linkage weight $\theta_{j,ik}$ with any unit j of U^A . In that case, the use of the estimation weight (4.2) underestimates the total Y^B . To solve this problem, the same solutions proposed in the context of the indicator variables $l_{j,ik}$ can be used. That is, a solution is to collapse two clusters in order to get at least one non-zero linkage weight $\theta_{j,ik}$. Unfortunately, this solution might require some manual intervention, which has been avoided up to now by not using the decision rule (2.3). A better solution is to impute a link by choosing one link at random within the cluster, and then assign arbitrarily a small value for $\theta_{j,ik}$ to the chosen link (for example, the smallest calculated non-zero linkage weight).

To estimate the total Y^B belonging to population U^B , one can use the estimator

$$\hat{Y}^{\text{RL}} = \sum_{i=1}^{n^{\text{RL}}} \sum_{k=1}^{M_i^B} w_{ik}^{\text{RL}} y_{ik}. \quad (4.3)$$

Following the same steps used to obtain equation (3.4), one can write \hat{Y}^{RL} as

$$\begin{aligned} \hat{Y}^{\text{RL}} &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^N \sum_{k=1}^{M_i^B} \theta_{j,ik} z_{ik}^{\text{RL}} \\ &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} z_j^{\text{RL}} \end{aligned} \quad (4.4)$$

where $z_{ik}^{\text{RL}} = Y_i / \Theta_{ik}$ for all $k \in U_i^B$, and $\Theta_i = \sum_{k=1}^{M_i^B} \Theta_{ik} = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} \theta_{j,ik}$. Using this last expression, it can be shown that \hat{Y}^{RL} is design unbiased for Y^B . The variance of \hat{Y}^{RL} is given by

$$\text{Var}(\hat{Y}^{\text{RL}}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j^{\text{RL}} Z_{j'}^{\text{RL}}. \quad (4.5)$$

4.2 Approach 2: Use all Non-Zero Links Above a given Threshold

Using all non-zero links with the GWSM as in Approach 1 might require the manipulation of large files of size $M^A \times M^B$. This is because it might turn out that most of the records between files A and B have non-zero linkage weights θ . In practice, even if this happens, we can expect that most of these linkage weights will be relatively small or negligible to the extent that, although non-zero, the links are very unlikely to be true links. In that case, it might be useful to only consider the links with a linkage weight θ above a given threshold θ_{High} .

For this second approach, we again no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not, but instead, we use the linkage weight $\theta_{j,ik}$ that are above the threshold θ_{High} . The linkage weights below the threshold are considered as zeros. We therefore define the linkage weight:

$$\theta_{j,ik}^T = \begin{cases} \theta_{j,ik} & \text{if } \theta_{j,ik} \geq \theta_{\text{High}} \\ 0 & \text{otherwise.} \end{cases}$$

For each unit j selected in s^A , we identify the units ik of U^B that have $\theta_{j,ik}^T > 0$. Let $\Omega^{\text{RLT},B}$ be the set of the n^{RLT} clusters identified by the units $j \in s^A$, where "RLT" stands for "Record Linkage with Threshold". Note that $n^{\text{RLT}} \leq n^{\text{RL}}$. On the other hand, we have $n^{\text{RLT}} = n$ if the record linkage between U^A and U^B is done by using the decision rule (2.3) with $\theta_{\text{High}} = \theta_{\text{Low}}$.

The initial weight $w_{ik}^{*\text{RLT}}$ is given by

$$w_{ik}^{*\text{RL}} = \sum_{j=1}^{M^A} \theta_{j,ik}^T \frac{t_j}{\pi_j^A}. \quad (4.6)$$

The final weight w_i^{RLT} is given by

$$w_i^{\text{RL}} = \frac{\sum_{k=1}^{M_i^B} w_{ik}^{*\text{RLT}}}{\sum_{k=1}^{M_i^B} \Theta_{ik}^T} \quad (4.7)$$

where $\Theta_{ik}^T = \sum_{j=1}^{M^A} \theta_{j,ik}^T$. Finally, we assign $w_{ik}^{\text{RLT}} = w_i^{\text{RLT}}$ for all $k \in U_i^B$. As for Approach 1, it is interesting to note that by letting $\alpha_{j,ik} = \theta_{j,ik}^T / \Theta_{ik}^T$ where $\Theta_{ik}^{T,B} = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik}^T$, we obtain, for the estimation weight w_i^{RLT} , an equivalent formulation to the one given by (3.7) and (3.8).

The number of zero linkage weights θ^T will be greater than or equal to the number of zero linkage weights θ used by Approach 1. Therefore, the constraint that each cluster i of U^B must have at least one non-zero linkage weight $\theta_{j,ik}^T$

with a unit j of U^A might be more difficult to satisfy. In that case, the use of the estimation weight (4.7) underestimate the total Y^B . To solve this problem, the same solutions proposed before can be used.

To estimate the total Y^B , one can use the same estimator as (4.3), where we replace the number of identified clusters n^{RL} by n^{RLT} , and the estimation weight w_{ik}^{RL} by w_{ik}^{RLT} . As for estimator (4.3), it can be shown that this estimator \hat{Y}^{RLT} is design unbiased.

4.3 Approach 3: Choose the Links by Random Selection

In order to avoid making a decision on whether there is a link or not between unit j from U^A and unit k of cluster i from U^B , one can decide to simply choose the links at random from the set of non-zero links. For this, it is reasonable to choose the links with probabilities proportional to the linkage weights θ . This can be achieved by Bernoulli trials where, for each pair (j, ik) , we decide on accepting a link or not by generating a random number $u_{j,ik} \sim U(0,1)$ that is compared to a quantity proportional to the linkage weight $\theta_{j,ik}$.

In the point of view of record linkage, this approach cannot be considered as optimal. When using the decision rule (2.3) of Fellegi and Sunter, the idea is to try to minimise the number false links and false nonlinks. The link $l_{j,ik}$ is accepted only if the linkage weight $\theta_{j,ik}$ is large enough (i.e., $\theta_{j,ik} \geq \theta_{\text{High}}$), or if it is moderately large (i.e., $\theta_{\text{Low}} < \theta_{j,ik} < \theta_{\text{High}}$) and has been accepted after manual resolution. Selecting the links randomly using Bernoulli trials might lead to the selection of links that would have not been accepted through the decision rule (2.3), even though the selection probabilities are proportional to the linkage weights. Some of the resulting links between the two populations U^A and U^B might then be false ones, and some units that are not linked might be false nonlinks. The linkage errors are therefore likely to be higher than if the decision rule (2.3) would be used. However, in the present context, the quality of the linkage is of secondary interest. The present problem is to try to estimate the total Y^B using the sample s^A selected from U^A , and not to evaluate the quality of the links. The precision of the estimates of Y^B will in fact be measured only in terms of the sampling variability of the estimators, by conditioning on the linkage weights $\theta_{j,ik}$. Note that this sampling variability will take into account the random selection of the links, but not the linkage errors.

The first step before performing the Bernoulli trials is to transform the linkage weights in order to restrict them to the $[0,1]$ interval. By looking at (2.1), it can be seen that the linkage weights $\theta_{j,ik}$ correspond in fact to a logit transformation (in base 2) of the probability $P(\mu_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})$. Similarly, the linkage weights given by (2.2) depend only on this probability. Hence, one way to transform the linkage weights is simply to use the

probability $P(\mu_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})$. From (2.1), we obtain this result by using the function $\tilde{\theta} = 2^{\tilde{\theta}}/(1 + 2^{\tilde{\theta}})$. From (2.2), we use $\tilde{\theta} = \theta/(1 + \theta)$. When the linkage weights are not obtained through (2.1) nor (2.2), a possible transformation is to divide each linkage weight by the maximum possible value $\theta_{\text{Max}} = \max_{j=1, i=1, k=1}^{M^A, N, M^B} \theta_{j,ik}$. Note that we assume that the linkages weights are all greater than or equal to zero, which is the case with definition (2.2), but not necessarily in general.

Once the adjusted linkage weights $\tilde{\theta}_{j,ik}$ have been obtained, for each pair (j, ik) , we generate a random number $u_{j,ik} \sim U(0,1)$. Then, we set the indicator variable θ_{Hig} to 1 if $u_{j,ik} \leq \tilde{\theta}_{j,ik}$, and 0 otherwise. This process provides a set of links similar to the ones used in the Classical Approach, with the exception that now the links have been determined randomly instead of through a decision process comparable to (2.3). Note that since $E(\tilde{l}_{j,ik}) = \tilde{\theta}_{j,ik}$, the sum of the adjusted linkage weights $\tilde{\theta}_{j,ik}$ corresponds to the expected total number of links L from the Bernoulli process in, $A \times B$, i.e.,

$$\sum_{j=1}^{M^A} \sum_{i=1}^N \sum_{k=1}^{M_i^B} \tilde{\theta}_{j,ik} = L. \quad (4.8)$$

For each unit j selected in s^A , we identify the units ik of U^B that have $\tilde{l}_{j,ik} = 1$. Let $\tilde{\Omega}^B$ be the set of the \tilde{n} clusters identified by the units $j \in s^A$. Note that $\tilde{n} \leq n^{\text{RL}}$. Unfortunately, in contrast to n^{RL} and n^{RLT} , the random number of clusters \tilde{n} is hardly comparable to n .

The initial weight \tilde{w}'_{ik} is defined as follows:

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^N \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j. \quad (4.9)$$

The final weight \tilde{w}_{ik} is given by

$$\tilde{w}_{ik} = \frac{\sum_{k=1}^{M_i^B} \tilde{w}'_{ik}}{\sum_{k=1}^{M_i^B} \tilde{l}_{ik}} \quad (4.10)$$

where $\tilde{l}_{ik} = \sum_{j=1}^{M^A} \tilde{l}_{j,ik}$. The quantity \tilde{l}_{ik} represents the realised number of links between the units of U^A and the unit k of cluster i of population U^B . Finally, we assign $\tilde{w}_{ik} = \tilde{w}_i$ for all $k \in U_i^B$.

To estimate the total Y^B , we can use the estimator

$$\hat{\hat{Y}} = \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{M_i^B} \tilde{w}_{ik} y_{ik}. \quad (4.11)$$

By conditioning on the accepted links \tilde{l} , it can be shown that estimator (4.11) is conditionally design unbiased and hence, unconditionally design unbiased. Note that by conditioning on \tilde{l} , the estimator (4.11) is then equivalent to

(3.1). To get the variance of $\hat{\hat{Y}}$, again conditional arguments need to be used. Letting the subscript 1 indicate that the expectation is taken over all possible sets of links, we have

$$\text{Var}(\hat{\hat{Y}}) = E_1 \text{Var}_2(\hat{\hat{Y}}) + \text{Var}_1 E_2(\hat{\hat{Y}}). \quad (4.12)$$

First, from conditional unbiasedness, we have

$$E_2(\hat{\hat{Y}}) = Y^B. \quad (4.13)$$

Therefore,

$$\text{Var}_1 E_2(\hat{\hat{Y}}) = 0. \quad (4.14)$$

Second, from (3.5), we directly have

$$\text{Var}_2(\hat{\hat{Y}}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{Z}_j \tilde{Z}_{j'}, \quad (4.15)$$

where \tilde{Z}_j is defined as in (3.4) but with the links l replaced by \tilde{l} . Hence, the variance of $\hat{\hat{Y}}$ can be expressed as

$$\text{Var}_2(\hat{\hat{Y}}) = E_1 \left(\sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{Z}_j \tilde{Z}_{j'} \right) \quad (4.16)$$

where the expectation is taken over all possible sets of links.

With the GWSM, we stated in section 3 a constraint that must be satisfied for unbiasedness of the GWSM. In the present approach, by randomly selecting the links, it is very likely that this constraint will not be satisfied. To solve this problem, we can impute a link by choosing the one with the highest non-zero linkage weight $\theta_{j,ik}$ within the cluster. If there is still no link because all $\theta_{j,ik} = 0$, it is possible to choose one link at random within the cluster. It should be noted that this solution preserves the design unbiasedness of the GWSM.

4.4 Some Remarks

The three proposed approaches do not use the decision rule (2.3). They also not make use of any manual resolution. Hence, the answer to the question (c) of the introduction is yes. That is, GWSM can help in reducing the manual resolution required by record linkage. Note that there is however a price to pay for avoiding manual resolution.

First, with Approach 1, the number n^{RL} of clusters identified by the units $j \in s^A$ is greater than or equal to the number n of clusters identified by the Classical Approach, i.e., when the decision rule (2.3) is used to identify the links. This is because we use all non-zero links, and not just the ones satisfying the decision rule (2.3). As a consequence, the collection costs with Approach 1 will be greater than or equal to the ones related to the use of the Classical Approach. It needs then to be checked which ones are the most important: the collections costs or the costs of manual resolution. Note that if the precision resulting from the use of Approach 1 is much higher than one from the Classical

Approach, it might be more of interest to use the former than the latter.

With Approach 2, we have $n^{\text{RLT}} \leq n^{\text{RL}}$ and therefore the collection costs of this approach are less than or equal to the ones of Approach 1. If the precision of Approach 2 is comparable to the one of Approach 1, then the former will certainly be more advantageous than the latter. By comparing Approach 2 with the Classical Approach, it can be seen that the collection costs can be almost equivalent if the value of the threshold θ_{High} is chosen to be close to the lower and upper thresholds of the decision rule (2.3). Note that Approach 2 is not using any manual resolution. If the precision of Approach 2 is at least comparable to the one of the Classical Approach, then Approach 2 will have a clear advantage. Note also that if $\theta_{\text{High}} = \theta_{\text{Low}}$, the two approach differs only in the definition of the estimation weights obtained by the GWSM. Approach 2 uses the linkage weights θ , while the Classical Approach uses the indicator variables l . After setting $\theta_{\text{High}} = \theta_{\text{Low}}$, it is certainly of interest to verify which approach has the highest precision.

With Approach 3, the number of selected links will be less than or equal to the number of non-zero links used by Approach 1, i.e., $\tilde{n} \leq n^{\text{RL}}$. Hence, the collection costs of Approach 3 will be less than or equal to the ones of Approach 1. In terms of precision, it is not clear which variance is likely to be the smallest between to two approaches. As mentioned before, in opposite to n^{RL} and n^{RLT} the random number of clusters \tilde{n} is hardly comparable to n . The two depends on different parameters: The Classical Approach depends on the thresholds θ_{Low} and θ_{High} , while Approach 3 depends on the adjusted linkage weights $\tilde{\theta}_{j,ik}$ that correspond to the selection probabilities of the links.

5. SIMULATION STUDY

A simulation study was performed to evaluate the proposed approaches against the Classical Approach where the decision rule (2.3) is used to determine the links. This study was made by comparing the precision obtained for the estimation of a total Y^B using five different approaches:

Approach 1: use all non-zero links with their respective linkage weights

Approach 2: use all non-zero links above a threshold

Approach 3: choose the links randomly using Bernoulli trials

Approach 4: Classical Approach

Approach 5: use all non-zero links, but with the indicator variable l

Approach 5 is a mixture of Approach 1 and the Classical Approach. It is basically to first accept as links all pairs (j, ik) with a non-zero linkage weights, i.e., assign $l_{j,ik} = 1$ for all pairs (j, ik) where $\theta_{j,ik} > 0$, and 0 otherwise. The GWSM described in section 3 is then used to produce the

estimate of Y^B . Approach 5 was added to the simulations to see the effect of using the indicator variable l instead of the linkage weight θ when using all non-zero links. As for the other approaches, Approach 5 can be shown to be unbiased.

Given that all five approaches yield design unbiased estimates of the total Y^B , the quantity of interest for comparing the various approaches was the standard error of the estimate, or simply the coefficient of variation (i.e., the ratio of the square root of the variance to the expected value).

The simulation study was performed based on the agriculture example mentioned throughout the paper. This example corresponds in fact to a real situation occurring at Statistics Canada related to the construction of the Whole Farm Data Base (see Statistics Canada 2000). Note that although the simulation study was based on a real situation, some of the numbers used have been changed for confidentiality reasons. Also, the linkage process did not reflect the exact procedure used within Statistics Canada. For more information on the exact procedure, see Lim (2000). It was felt that these changes do not negate the results of the simulation study. The main purpose of the simulations was to evaluate the proposed approaches against the Classical Approach. It was not intended to solve the problems related to the construction of the Whole Farm Data Base, which could be considered as a secondary goal.

Recall that the agriculture example is one of an agricultural survey where the first population U^A is a list of farms as determined by the Canadian Census of Agriculture. This list is from the 1996 Farm Register, which is essentially a list of all records collected during the 1991 Census of Agriculture with all the updates that have occurred since 1991. It contains a farm operator identifier together with some socio-demographic variables related to the farm operators. The second population U^B is a list of taxation records from the CCRA. This second list is the 1996 Unincorporated CCRA Tax File that contains data on tax filers declaring at least one farming income. It contains a household identifier (only on a sample basis), a tax filer identifier, and also socio-demographic variables related to the tax filers.

At Statistics Canada, Agriculture Division produces estimates on crops and livestock using samples selected from the Farm Register (population U^A). To create the Whole Farm Data Base, it is of interest to collect tax data for the farms that have been selected in the samples from the Farm Register. This is done by first merging the Farm Register with the Unincorporated CCRA Tax File (population U^B) and then obtaining the tax data from CCRA. As mentioned before, it turns out that the relationship between the farm operators of the Farm Register and the tax filers from the Unincorporated CCRA Tax File is not one-to-one. This is why the GWSM turns out to be a useful approach for producing estimation weights for the tax filers selected through the sample of farm operators from the Farm Register.

Some might argue that there is no need to obtain a set of clusters identified by the units $j \in s^A$, since the target population U^B is one of tax filers from the Unincorporated CCRA Tax File, which is usually available on a census basis. Note however that this is not totally true. Not all variables of interest are available on this file and Statistics Canada needs to pay for the extra variables requested from CCRA. Also, the data from the Unincorporated CCRA Tax File are not free of errors due to keying, coding, *etc.*, and therefore there are some costs related to cleaning up the data. For these reasons, it is found preferable to restrict the data from the target population U^B to a subset only. Since this needs to be done, one way of identifying the set of clusters to be used in the estimate of Y^B is simply to do it through the sample s^A selected from U^A .

Apart from the Classical Approach, all approaches consider the linkage itself between U^A and U^B as a secondary goal, the first one being to produce an estimate Y^B for the target population U^B . However, the application mentioned here is one related to the Whole Farm Data Base, which aims to be an integrated data base. Not having a linkage of good quality between the populations U^A and U^B would lead to erroneous microdata analyses between the crops and livestock variables measured in the sample s^A and the tax data obtained from U^B . On this aspect, the authors agree that the proposed approaches, with the exception of the Classical Approach, are not viable in the present context. This is true however in a long term point of view. Because manual resolution is needed when using a decision rule such as (2.3), one could suggest to use the proposed approaches to produce some of the required estimates from U^B in the short term, before the final linkage is available, after manual resolution. Recall that the main purpose of the simulations is to evaluate the proposed approaches against the Classical Approach. The agriculture example has not been chosen because it corresponds to a real situation, but more because of the availability of the data. It could have been any other example such as the other one mentioned in the introduction where U^A is a population of parents and U^B a population of children belonging to the parents.

For the purpose of the simulations, two provinces of Canada were considered: New Brunswick and Québec. The former can be considered as a small province and the latter a large one. Table 1 provides the size of the different files. Because the household identifier is not available for the entire population U^B , for the purpose of the simulations, it has been constructed based on a sample. This sample has the household identifier coded for each tax filer. For the non-sample tax filers, the household identifiers were randomly assigned such that the household sizes correspond to the same proportions of household sizes found in the sample.

Table 1

Agriculture Example

	Québec	New Brunswick
Size of Farm Register (U^A)	43017	4930
Size of Tax File (U^B)	52394	5155
Total number of households of U^B	22387	2194
Total number of Non-zero Linkage Weights	105113	13787

The linkage process used for the simulations was a match using five variables. It was performed using the MERGE statement in SAS®. All records on both files were compared to one another in order to see if a potential match had occurred. The record linkage was performed using the following five key variables common to both sources:

- first name (modified using NYSIIS)
- last name (modified using NYSIIS)
- birth date
- street address
- postal code

The first name and last name variables were modified using the NYSIIS system. This basically changes the name in phonetic expressions, which in turn increases the chance of finding matches by reducing the probability that a good match is rejected because of a spelling mistake. For more details about NYSIIS, see Lynch and Arends (1977).

Records that matched on all 5 variables received the highest linkage weight ($\theta=60$). Records that matched on only a subset of at least 2 of the 5 variables received a lower linkage weight (as low as $\theta=2$). It should be noted that the levels of the linkage weights were chosen arbitrarily. As mentioned before, it is not really the levels themselves that are important, but rather the relative importance of the linkage weights between each other.

Records that did not match on any combination of key variables were not considered as potential links, which is equivalent as having a linkage weight of zero. Two different thresholds were used for the simulations: $\theta_{\text{High}} = \theta_{\text{Low}} = 15$ and $\theta_{\text{High}} = \theta_{\text{Low}} = 30$. The upper and lower thresholds, θ_{High} and θ_{Low} , were set to be the same to avoid the grey area where some manual intervention is needed when applying the decision rule (2.3).

Note that the constraint related to the use of the GWSM needed to be satisfied. When for a cluster i of U^B there was no non-zero linkage weight $\theta_{j,ik}$ between any units k of this cluster and the units from U^A , we imputed a link by choosing the link with the largest linkage weight $\theta_{j,ik}$ within the cluster. Note that it also happened that for some units j of U^A , there was no non-zero linkage weight $\theta_{j,ik}$ with any unit ik of U^B , this was not considered a problem since the only coverage in which we are interested is the one of U^B . Table 1 provides the total number of non-zero links found in each of the two provinces.

For the simulations, we have selected the sample from U^A (*i.e.*, the Farm Register) using Simple Random Sampling Without Replacement (SRSWOR), without any stratification. We also considered two sampling fractions: 30% and 70%. The quantity of interest Y^B to be estimated was the Total Farming Income.

Since we have the whole population of farms and taxation records, it was possible for us to calculate the theoretical variance for these estimates. It was also possible to estimate this variance by selecting a large number of samples (*i.e.*, performing a Monte-Carlo study), estimating the parameter Y^B for each sample, and then calculating the variance of all the estimates. Both approaches were used. For the simulations, 500 simple random samples were selected for each approach for the two different sampling fractions (30% and 70%). The two thresholds (15 and 30) were also used to better understand the properties of the given estimators.

Because we assumed SRSWOR, the theoretical formulas given in section 4 could be simplified. For example, under SRSWOR, the variance formula (4.5) reduced to the following:

$$\text{Var}(\hat{Y}^{\text{RL}}) = M^A \frac{(1-f)}{f} S_{Z, \text{RL}}^2 \quad (5.1)$$

where $f = m^A/M^A$ is the sampling fraction, $S_{Z, \text{RL}}^2 = 1/M^A - 1 \sum_{j=1}^{M^A} (Z_j^{\text{RL}} - \bar{Z}^{\text{RL}})^2$ and $\bar{Z}^{\text{RL}} = 1/M^A \sum_{j=1}^{M^A} Z_j^{\text{RL}}$.

The Monte-Carlo study involved 500 replicates. For each of the two sampling fractions (30% and 70%), 500 simple random samples t were selected, and the expectation and variance for each of the five approaches were then estimated using

$$\hat{E}(\hat{Y}) = \frac{1}{500} \sum_{t=1}^{500} \hat{Y}_t \quad (5.2)$$

and

$$\hat{V}(\hat{Y}) = \frac{1}{500} \sum_{t=1}^{500} (\hat{Y}_t - \hat{E}(\hat{Y}))^2. \quad (5.3)$$

The estimated coefficients of variation (CVs) were obtained by using

$$C\hat{V}(\hat{Y}) = 100 \times \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{E}(\hat{Y})}. \quad (5.4)$$

The Monte-Carlo process was performed to verify empirically the exactness of the theoretical formulas provided in section 4. The results indicate that all the theoretical formulas provided were exact.

The results of the study are presented in Figures 2.1 to 2.4, Table 2, and Figure 3. Figures 2.1 to 2.4 provide bar charts of the CVs obtained for each of the five approaches. The bar charts are given for the eight cases obtained by crossing the two provinces Québec and New Brunswick, the two sampling fractions 30% and 70%, and the two thresholds 15 and 30. On each bar of the charts, one can find the number of non-zero links between U^A and U^B for

each of the five approaches. Note that for Approach 3, it corresponds in fact to the expected number of non-zero links. The number of (expected) non-zero links does not change from one sampling fraction to another. Table 2 shows the average number of clusters interviewed by approach, for each of the eight cases, where the average is taken over the 500 samples used for the simulations. The numbers in parenthesis are the standard deviations. They are relatively small compared to the averages and therefore the number of clusters identified through the sample s^A does not fluctuate greatly from one sample s^A to another. Figure 3 provides scattered plots of the obtained CVs by the average number of clusters identified through the sample s^A , for each of the eight cases.

By looking at the Figures 2.1 to 2.4, it can be seen that in all cases, Approach 1 and Approach 5 provided the smallest CVs for the estimation of the Total Farming Income. Therefore, using all non-zero links yield the greatest precision. Note however that by looking at Table 2, we can see that these approaches also lead to the highest number of clusters identified through the sample selected from U^A . In fact, we can see that the greater the number of clusters used in the estimation is, the greater the precision of the resulting estimates is. This result is shown in Figure 3 where we can see that the CVs tend to decrease as the average number of clusters identified through s^A increases. Although this result is well known in the classical sampling theory, it was not guaranteed to hold in the context of the GWSM. As we can see from equation (3.5), it is not the sample size of s^A that increases, but rather the homogeneity of the derived variables Z_j .

Now, by comparing Approach 1 and Approach 5, it can be seen that the latter always provided the smallest variance. Therefore, this suggests to use the indicator variable l instead of the linkage weight θ when using all non-zero links. Note that it seems this can be generalised since the same phenomenon occurred with Approach 2 and Approach 4 (Classical Approach). Recall that, because $\theta_{\text{High}} = \theta_{\text{Low}}$, the two approaches differ only in the definition of the estimation weights obtained by the GWSM. Approach 2 uses the linkage weights θ , while the Classical Approach uses the indicator variables l . Note that this results goes along the conclusions of Kalton and Brick (1995) since the optimal choice of letting the constants α being 0 for some units and a positive value that is equal for all the remaining units within the cluster corresponds to the use of the indicator variable l .

We now concentrate on Approach 3. For seven out of the eight histograms of Figures 2.1 to 2.4, Approach 3 produced the highest CVs. The only lower CV was obtained for Québec, with the sampling fraction of 30% and the threshold $\theta_{\text{High}} = 30$. It should however be noted that this approach is the one that used the lowest number of non-zero links, and also the lowest average number of clusters identified through s^A . Therefore, this result is not totally surprising. Recall that the number of non-zero links

used by Approach 3 does not depend on the threshold θ_{High} and thus the CVs obtained for Québec with $f=0.3$ were equal for $\theta_{\text{High}}=15$ and $\theta_{\text{High}}=30$. For $\theta_{\text{High}}=15$, the CV obtained for Approach 3 for Québec was higher than the ones for Approaches 2 and 4, and these two were using more non-zero links, and more clusters. For $\theta_{\text{High}}=30$ the CV obtained for Approach 3 was lower than the ones from approaches 2 and 4, but these two were still using more non-zero links, and more clusters. Therefore, there are intermediate situations where with $15 < \theta_{\text{High}} < 30$, we should get equal CVs for approaches 3 and 2, and approaches 3 and 4. As a consequence, to get equal CVs between Approach 3 and each of approaches 2 and 4, more non-zero links and more clusters must be used by the latter. This suggests that in some cases, Approach 3 might be more appropriate to use than approaches 2 and 4 because estimates with the same precision can be obtained with lower collection costs.

In order to better compare Approach 3 to the approaches 2 and 4, we forced the number of expected non-zero links to be the same as the number of non-zero links used by approaches 2 and 4. For this, we have transformed the linkage weights $\theta_{j,ik}$ to $\tilde{\theta}_{j,ik}$ in order to have

$$\sum_{j=1}^{M^A} \sum_{i=1}^N \sum_{k=1}^{M^B} \tilde{\theta}_{j,ik} = L_0 \tag{5.5}$$

where L_0 is the desired number of non-zero links. The transformation used was

$$\tilde{\theta}_{j,ik} = \begin{cases} \theta_{j,ik} / \theta_* & \text{if } \frac{\theta_{j,ik}}{\theta_*} \leq 1 \\ 1 & \text{otherwise} \end{cases} \tag{5.6}$$

where θ_* was determined iteratively such that (5.5) is satisfied. The use of Approach 3 with the transformation (5.6) is referred to as Approach 6. The results of the simulations are presented in Figures 4.1 to 4.4. As we can see, Approach 6 turned out to have the smallest CVs for half of the cases. For the other cases, Approach 4 yielded the best precision. Note that this situation did not occur for a particular province only, nor a particular sampling fraction, and also nor for a particular threshold. It would therefore be difficult in practice to determine in advance which of Approach 6 or Approach 4 would produce the smallest CVs. Because of this, and because of the fact that Approach 6 (and Approach 3) can produce large linkage errors, Approach 4 should be preferred.

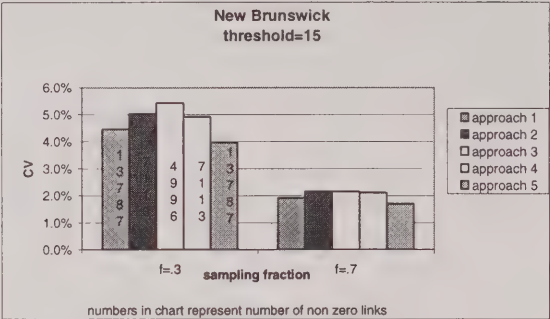


Figure 2.1 CVs for New Brunswick (with $\theta_{\text{High}} = \theta_{\text{Low}} = 15$.)

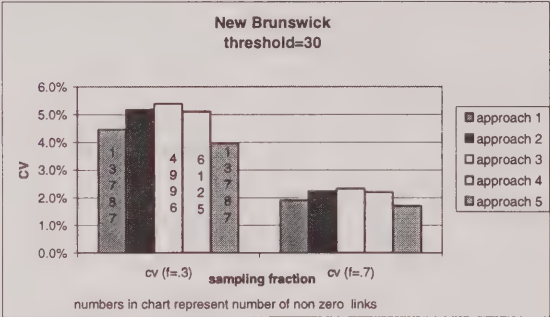


Figure 2.2 CVs for New Brunswick (with $\theta_{\text{High}} = \theta_{\text{Low}} = 30$)

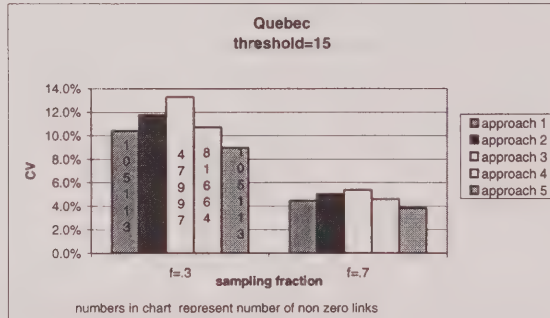


Figure 2.3 CVs for Québec (with $\theta_{\text{High}} = \theta_{\text{Low}} = 15$)

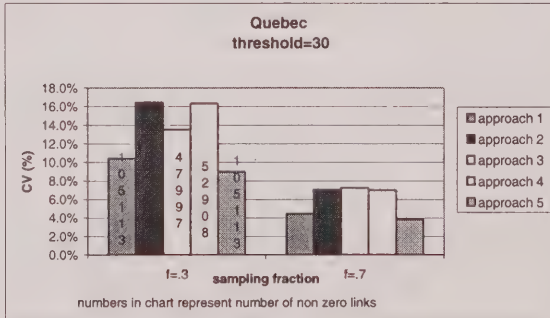


Figure 2.4 CVs for Québec (with $\theta_{\text{High}} = \theta_{\text{Low}} = 30$)

Table 2
Average Number of Identified Cluster

Threshold	Approach	Average number of identified clusters (s.e.)			
		Quebec		New Brunswick	
		f=.3	f=.7	f=.3	f=.7
15	1	15752(58)	21106(30)	1709(18)	2100(7)
	2	14281(49)	20593(34)	1310(17)	1966(13)
	3	10930(50)	18881(47)	1123(14)	1869(14)
	4	14281(49)	20593(34)	1310(17)	1966(13)
	5	15752(58)	21106(30)	1709(18)	2100(7)
30	1	15752(58)	21106(30)	1709(18)	2100(7)
	2	11310(45)	19139(37)	1215(17)	1924(15)
	3	10930(50)	18881(47)	1123(14)	1869(14)
	4	11310(45)	19139(37)	1215(17)	1924(15)
	5	15752(58)	21106(30)	1709(18)	2100(7)



Figure 3. Graphs of CVs versus Average Number of Identified Clusters

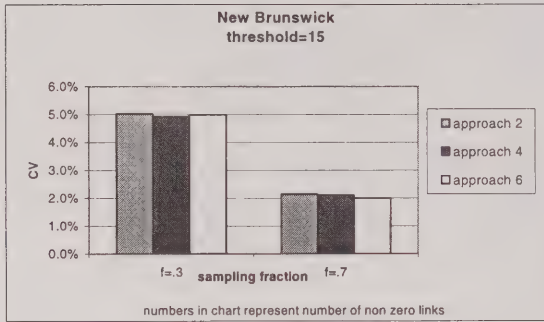


Figure 4.1. CVs for New Brunswick (with $\theta_{\text{High}} = \theta_{\text{Low}} = 15$).

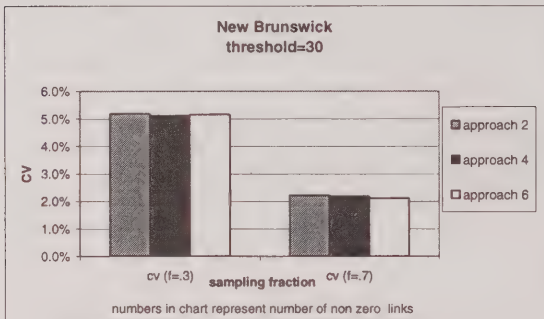


Figure 4.2. CVs for Québec (with $\theta_{\text{High}} = \theta_{\text{Low}} = 30$).

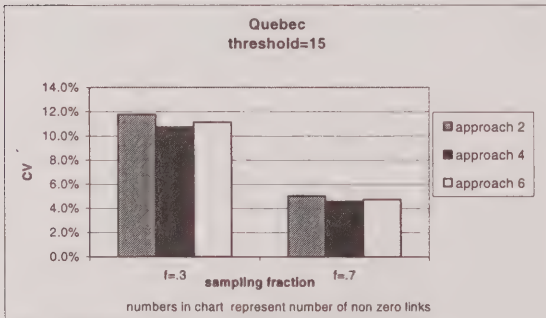


Figure 4.3. CVs for Québec (with $\theta_{\text{High}} = \theta_{\text{Low}} = 15$).

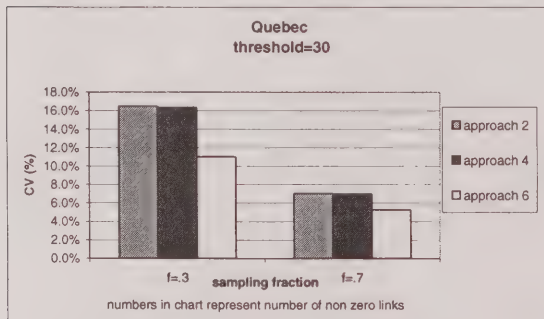


Figure 4.4. CVs for Québec (with $\theta_{\text{High}} = \theta_{\text{Low}} = 30$).

6. CONCLUSION

In the present paper, we have seen that the GWSM is adaptable to populations linked through Record Linkage. This is in fact simply a natural extension of the case where the links are either present or absent, which corresponds to the use of an indicator variable $l_{j,ik} = 1$ if the pair (j, ik) is considered to be a link, 0 otherwise. When two populations are linked through record linkage, there is always some uncertainty left because the decisions on the links are made using a probabilistic approach. Therefore, replacing the indicator variable $l_{j,ik}$ by the linkage weight $\theta_{j,ik}$ that has been computed for each pair (j, ik) simply makes the GWSM more generalised.

Some simulations were performed using the 1996 Farm Register (population U^A) and the 1996 Unincorporated CCRA Tax File (population U^B). We compared the variances obtained for each of the five approaches: (1) use all non-zero links; (2) use all non-zero links above a threshold; (3) choose links randomly using Bernoulli trials; (4) Classical Approach; (5) use all non-zero links, but with the indicator variable l . All results showed that Approach 1 and Approach 5 provide the smallest CVs for the estimation of the Total Farming Income. These two approaches use however the highest number of links, and also the highest number of clusters identified through s^A , which implies the highest collection costs. Because of this, the approaches 2, 3 and 4 might be viewed as good compromises.

For a given threshold θ_{High} , it is preferable to use the indicator variable l instead of the linkage weights θ in the construction of the estimation weights with the GWSM. This result holds even for $\theta_{\text{High}} = 0$ (i.e., no threshold is used), as for approaches 1 and 5. The estimates produced with the indicator variable l always had the smallest CVs and this result goes along the conclusions of Kalton and Brick (1995). Hence, Approach 5 should be preferred to Approach 1, and Approach 4 should be preferred to Approach 2.

The use of the threshold θ_{High} is useful to reduce the number of non-zero links to be manipulated. By reducing the number of non-zero links, we reduce as well the number of clusters identified through the sample s^A , and hence we reduce the collection costs associated to the measurement of the variable of interest y within the clusters. Note that by reducing the number of links, we decrease the precision of the estimates produced. Therefore, a choice needs to be made between the desired precision and the collection costs.

The reduction of the number of non-zero links can also be achieved by using the decision rule (2.3) with the two thresholds θ_{Low} and θ_{High} . This decreases the collection costs, but introduces the need of some manual resolution when the linkage weights θ are between θ_{Low} and θ_{High} . The manual resolution leads however to better links, i.e., with less linkage errors. If manual resolution is used only to make the links one-to-one between population U^A and

population U^B , then it might not be necessary since the GWSM is particularly appropriate to handle estimation in situations where the links between U^A and U^B are complex.

When compared to approaches 2 and 4, Approach 3 turned out to be preferable in some cases. Because it would be difficult in practice to determine in advance which of Approach 3 or Approach 4 would produce the smallest CVs, and because of the fact that Approach 3 can produce large linkage errors, Approach 4 should be preferred. Hence, the Classical Approach of using the GWSM with the indicator variable l with links determined using a decision rule such as (2.3) seems the most appropriate approach to estimate the total Y^B using a sample selected from U^A .

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and the two referees for their useful suggestions and comments. These have contributed to improve significantly the quality of the paper.

REFERENCES

- BARTLETT, S., KREWSKI, D., WANG, Y. and ZIELINSKI, J.M. (1993). Evaluation of error rates in large scale computerized record linkage studies. *Survey Methodology*, 19, 3-12.
- BELIN, T.R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology*, 19, 13-29.
- BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.
- BUDD, E.C., and RADNER, D.B. (1969). The OBE size distributions series: methods and tentative results for 1964. *American Economic Review*, Papers and Proceedings, LIX, 435-449.
- ERNST, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.
- FELLEGI, I.P., and SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- GAILLY, B., and LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg.
- KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LIM, A. (2000). Results of the Linkage between the 1998 Taxation Data and the 1998 Farm Register. Internal document of the Business Survey Methods Division, Statistics Canada.
- LYNCH, B.T., and ARENDS, W.L. (1977). Selection of a Surname Coding Procedure for the SRS Record Linkage System. Document of the Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
- NEWCOME, H.B., KENNEDY, J.M., AXFORD, S.J. and JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.
- OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.
- SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SINGH, A.C., MANTEL, A.J., KINACK, M.D. and ROWE, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.
- STATISTICS CANADA (2000). *Whole Farm Database reference manual*. Publication No. 21F0005GIE, Statistics Canada, 100 pages.
- THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.
- THOMPSON, S.K., and SEBER, G.A. (1996). *Adaptive Sampling*. New York: John Wiley and Sons.
- WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley and Sons, 355-384.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.

Cross-sectional Estimation in Multiple-Panel Household Surveys

TAKIS MERKOURIS¹

ABSTRACT

This paper presents weighting procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation. The non static nature of a repeated panel survey is discussed in relation to estimation of population parameters at any wave of the survey. A repeated panel survey with overlapping panels is described as a special type of multiple frame survey, with the frames of the panels forming a time sequence. The paper proposes weighting strategies suitable for various multiple-panel survey situations. The proposed weighting schemes involve an adjustment of weights in domains of the combined panel sample that represent identical time periods covered by the individual panels. A weight adjustment procedure that deals with changes in the panels over time is discussed. The integration of the various weight adjustments required for cross-sectional estimation in a repeated panel household survey is also discussed.

KEY WORDS: Repeated panel surveys; Multiple frames; Temporal domains; Combined panels; Cross-sectional weighting; Weight share method.

1. INTRODUCTION

A panel survey collects the survey data for the same sample elements at different time points (the survey waves). A repeated panel survey is made up of a series of panel surveys, each having fixed duration, with the panels selected at different time points. In a repeated panel household survey a sample of households is selected for each panel from the population of households existing at the start of the panel. Depending on the objectives of the panel survey, one or all individuals in the sampled households become panel members to be followed throughout the duration of the panel or until they leave the survey population. At a subsequent survey wave the household sample consists of all the households in which panel members reside. A review of various types of panel surveys is given in Kalton and Citro (1993). A formalization of related concepts can be found in Deville (1998).

The type of repeated panel household survey considered in this paper consists of two or more panels covering overlapping time periods. A typical example of such a survey is the Canadian Survey of Labour and Income Dynamics (SLID), which employs two overlapping panels of duration of six years each; for a description of the SLID see Lavigne and Michaud (1998). In the SLID, each new panel is introduced three years after the introduction of the previous one. The sample for each panel is made up of two rotation groups from the Canadian Labour Force Survey, which uses a stratified multistage design with an area frame wherein dwellings containing households are the final sampling units.

A panel survey, though primarily conducted for longitudinal purposes, may also be used to produce cross-sectional estimates of population parameters for any survey wave. For cross-sectional purposes, data are usually collected at each survey wave for all individuals living in

households that contain at least one selected member. The process of obtaining cross-sectional estimates at any wave of a panel household survey after the first wave presents difficulties arising from the population and panel dynamics. Weighting schemes that deal with dynamic features of a single panel, such as movers and "cohabitants," have been discussed in the literature; see Kalton and Brick (1995), and Lavallée (1995) for details. Yet, there seems to be a paucity of work in the literature on cross-sectional estimation for repeated panel household surveys with overlapping panels; some initial work in the context of the SLID can be found in Lavallée (1994). The cross-sectional estimation problem in such multiple panel surveys is a proper combination of the panels that would account for the changes in the population and in the panels over time.

This paper describes procedures for cross-sectional estimation that combine information from overlapping panels of a repeated panel household survey. The coverage of the population by the individual panels at any given wave, and the use of the combined panels supplemented by a "top-up" sample to construct a representative cross-sectional sample are discussed in section 2. Also discussed in the same section are analogies with a multiple-frame survey scheme, as well as issues related to the sample dynamics. The weighting and estimation problem in repeated panel household surveys is described in section 3. Weighting strategies suitable for various panel survey situations are then proposed. Bias and efficiency issues related to the combination of panels are discussed. A weight adjustment procedure that deals effectively with changes in the combined panels over time is described in section 4. The integration of the various weight adjustments required for cross-sectional estimation in a repeated panel household survey is discussed in section 5. Finally, a summary and concluding remarks are provided in section 6.

¹ Takis Merkouris, Statistics Canada, Household Survey Methods Division, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

2. GENERAL CONSIDERATIONS

2.1 Coverage of the Cross-sectional Population

Important to cross-sectional estimation are changes in the population composition over time, occurring when individuals leave or enter the population. In a single-panel household survey, new entrants who have joined the survey population since the start of the panel are not represented in the sample at later waves if they live in households that do not contain any members of the original population. A multiple-panel household survey with overlapping panels provides a better coverage of the survey population than a single-panel survey, as it reduces the time period not covered by any of the panels. In the case of the SLID, this time period is reduced from a maximum of six years to a maximum of three years. Nevertheless, the problem of complete coverage remains unless a special supplementary sample of the non-covered population is taken at each survey wave. A survey scheme involving one panel and a supplementary sample drawn at each survey wave for cross-sectional purposes is described in Lavallée (1995). An alternative approach involves the selection, at each wave, of a new sample that covers the entire survey population but does not form a new panel. This sample (henceforth to be called top-up sample) is to be used only once, for cross-sectional purposes, and its size would normally be smaller than a panel's size. In the context of constructing a cross-sectional sample, a top-up sample is discussed as a non-trivial case of supplementary sample, essentially treated as an additional small overlapping panel.

The situation with regard to individuals who leave the population is as follows. For any panel, the sampling frame for the survey population at a time point t is essentially the sampling frame for the population at the start of the panel, with the leavers in the intervening period being treated as blanks on the frame. Panel members who leave the population before time t correspond to blanks on the frame, and thus their effect on cross-sectional estimates at time t is loss of efficiency but not bias; see also Kalton and Brick (1995) for relevant discussion.

The foregoing observations lead to the following perspective regarding the coverage of the population by each of the panels at any wave of the survey. As regards cross-sectional representation, each panel covers at the time of its selection the entire survey population represented by the preceding panels. Accordingly, the frames of the panels form a time sequence, with the frame of each panel containing at the start of the panel the frames of the preceding panels. In such a sequence of frames, a common frame is formed sequentially as the intersection of the frame of a new panel with the remainder of the original common frame of the preceding active panels. At any wave the common frame is the common frame at the start of the most recent of these panels, but without the leavers. The non-overlap frame domain at the start of a new panel consists of

individuals who entered the population after the start of the preceding panel. Other frame domains (relatively very small in size) may be formed by returning units of older frames, in which case the time sequence of frames is not completely nested. Because of the latter type of frame domains, the complete frame at any wave after the selection of the most recent panel is the union of the frames of all panels at that time point, not just the remainder of the frame of the most recent panel. In panel surveys which employ a top-up sample at each wave the complete frame is that of the top-up sample.

2.2 A Multiple Frame Analogy

With the above considerations, a multiple panel survey with overlapping panels can be thought of as a special type of multiple frame survey, in which the frame for the cross-sectional population is the union of mutually exclusive temporal domains defined by the frames of the panels and their intersections. The sizes of the frames of the individual panels, as well as the characteristics of the population members in each panel's frame, change over time. This is in contrast with the static character of the usual type of multiple frame survey. Also, there is a high degree of nesting in the sequence of panel frames, so that the total number of mutually exclusive temporal frame domains is small. Among the various frame domains the one that is common to all panels is by far the largest. These special multiple frame features have implications in cross-sectional estimation, as will be discussed in the next section.

The sample temporal domains may be even less static because of attrition, moves of selected individuals within and between panels and moves of non-selected individuals into households in which panel members reside. For instance, with the presence of new entrants (*e.g.*, immigrants) in households that contain selected individuals, a panel crosses the boundary of its frame into the frame of the succeeding panel.

The analogy with multiple-frame survey sampling places the problem of cross-sectional estimation for repeated surveys with overlapping panels into a familiar framework. However, the distinctive dynamic features of multiple panel surveys will have to be considered if conventional multiple frame approaches are contemplated for the formulation of a cross-sectional estimation methodology.

For the purpose of introducing a cross-sectional estimation procedure that combines information from the panels of a repeated panel household survey, it suffices to consider the simple situation involving two overlapping panels at the time point of the start of the second panel. Note that this would always be the situation in a survey with one panel and a top-up sample. Thus, adopting standard multiple frame notation, with B and A denoting the frames of the first and the second panel ($B \subset A$) at the start of the second panel, and with s_B , s_A denoting the respective samples, the setting can be presented schematically as in Figure 1.

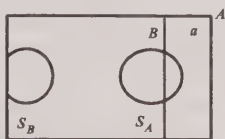


Figure 1. Two overlapping panels at the start of the second panel.

In Figure 1, A is the complete frame, so that the second panel at its start represents the cross-sectional population at that time. The overlap domain B is the remainder of the original frame of the first panel. The domain $a = B^c \cap A$ consists of all new entrants into the population since the start of the first panel. The samples s_B and s_A are the originally selected ones, with s_B reduced in size because of leavers and non respondents. It is assumed that the samples s_A and s_B are drawn independently from A and B according to specified probability designs $p_A(s_A)$ and $p_B(s_B)$, which determine the inclusion probabilities π_{Ai} and π_{Bi} of the i -th unit (household or any individual within it) for the original samples s_A and s_B , respectively. The samples s_A and s_B may intersect, since members in the overlap frame B can be selected in both panels. The issue of panel (sample) overlap is akin to that of duplicate sample units in multiple frame surveys. In repeated panel household surveys an operational constraint motivated by respondent burden may be to exclude from s_A individuals already selected in s_B , thus inducing $s_A \cap s_B = \emptyset$; for a discussion on this see Lavallée (1994). Here, as in any multiple frame situation, it is observed that if the probabilities π_{Ai} and π_{Bi} are small the probability of duplicate units is negligible. It will be assumed in the following that the probabilities π_{Ai} and π_{Bi} are small, and in effect $s_A \cap s_B = \emptyset$.

3. CROSS-SECTIONAL WEIGHTING AND ESTIMATION

This section describes procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation of population parameters. The discussion is confined to estimation of totals. A uniform approach to cross-sectional estimation for households and individuals is presented. This approach is based on the production of a set of weights for the combined panel sample that yield design-unbiased estimators of cross-sectional totals. Essentially, it involves the construction of a combined cross-sectional sample by means of an adjustment of the sampling weights of units from the temporal domains of the different panels that represent identical temporal domains of the cross-sectional population. While the delineation of the various temporal frame domains is necessary for determining the coverage of parts of the cross-sectional population by the different panels, the identification of some of the corresponding sample domains

may not be possible under the operating procedures of a repeated panel household survey. For example, the information needed to determine whether or not a unit in the second panel belongs to the non-overlap frame domain a (see Figure 1) may not be available. In this section, both cases of identifiable and non-identifiable temporal sample domains are considered. The weight adjustment for the combination of the panels involves only sampled units, and takes no account of any changes (other than leavers) in household membership between waves. A "weight share" adjustment that handles such changes should follow the combination of the panels, as it can be applied readily only to the combined sample; see relevant discussion in section 4.

3.1 Identifiable Temporal Sample Domains

Weighting options for the combination of the panels

For the construction of a cross-sectionally representative combined sample, a panel survey scheme such as that depicted in Figure 1 is considered. In analogy with a standard multiple frame argument (Bankier 1986; Skinner and Rao 1996) the two samples s_A and s_B can be thought of as selected independently from the complete frame A according to the sampling designs $p_A(s_A)$ and $p_B(s_B)$, but with a fixed time lag between the two selections. Then the two sampling designs $p_A(s_A)$ and $p_B(s_B)$ induce a well-defined design $p(s)$ on the set of samples $s = s_A \cup s_B$ in A . Thus conventional estimators, based on a single frame and a combined sample, may be constructed from $p(s)$. The standard approach, leading to the Horvitz-Thompson estimator, would be to assign sample units weights made inversely proportional to their inclusion probabilities. The probability of inclusion of the i -th population unit in the combined sample, $\pi_i = P(i \in s)$, is equal to $\pi_{Ai} + \pi_{Bi} - \pi_{Ai} \pi_{Bi}$ if $i \in B$, and equal to π_{Ai} if $i \in a$. The weight of the i -th unit of the sample is then $w_i = 1/\pi_i$. This weighting scheme can be used provided that it is possible to identify the common units in the samples s_A and s_B , so that the duplicate units can be eliminated. A simpler approach, especially for surveys with more than two panels, would be to assign any unit $i \in B$ a weight made inversely proportional to the expected number of selections of the unit, that is, inversely proportional to $\pi_{Ai} + \pi_{Bi}$. This weighting scheme, proposed by Kalton and Anderson (1986) for multiple frame surveys, does not require identification of duplicate sample units. Now, consider the sample domains $s_{ab} = s_A \cap B$ and $s_a = s_A \cap a$ of s_A . Also, let a value y_i be associated with population unit i for some population characteristic, and define the population total $Y_A = \sum_A y_i$ ($= \sum_B y_i + \sum_a y_i$). Then, employing the latter weighting scheme the unbiased estimator

$$\hat{Y}_A = \sum_s w_i y_i = \sum_{s_B} (\pi_{Ai} + \pi_{Bi})^{-1} y_i + \sum_{s_{ab}} (\pi_{Ai} + \pi_{Bi})^{-1} y_i + \sum_{s_a} \pi_{Ai}^{-1} y_i \quad (1)$$

of the total Y_A can be constructed. On the assumption that the probabilities π_{Ai} and π_{Bi} for $i \in s \cap B$ are small, the estimator \hat{Y}_A is approximately equal to the Horvitz-Thompson estimator.

The approach leading to the estimator (1) is not in general feasible, since the determination of the weight $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$ for $i \in s \cap B$ requires knowledge of π_{Ai} for units in s_B , and knowledge of π_{Bi} for units in s_{ab} . This is difficult or impossible to ascertain in household surveys because of stratified multistage sampling. In multiple-panel household surveys additional complications arise from the time element. For units that move (e.g., to another stratum) in the time between the selection of the panels it is impossible to determine both π_{Ai} and π_{Bi} .

An alternative strategy needs to be considered for developing weights for the sample overlap domain $s \cap B$. One approach that provides a general framework for handling this problem requires information on the probability of inclusion in only one of s_A or s_B , thus avoiding the difficulty noted above. The essence of the alternative approach considered here is to associate with the i -th unit from the overlap frame B a number p_i ($0 \leq p_i \leq 1$) when the unit is selected in s_B , and the number $1 - p_i$ when the unit is selected in s_A , and then define the weight of the unit as

$$w_i^* = p_i \frac{1}{\pi_{Bi}} I\{i \in s_B\} + (1 - p_i) \frac{1}{\pi_{Ai}} I\{i \in s_{ab}\}, \quad i \in B, \quad (2)$$

where I is the usual sample membership indicator variable. Clearly, $E(w_i^*) = 1$ under $p(s)$, and thus the use of the weights w_i^* will yield unbiased estimators $\hat{Y}_B = \sum_B w_i^* y_i$ for the total $Y_B = \sum_B y_i$, for any choice of constants p_i satisfying $0 \leq p_i \leq 1$, and for any sampling designs $p_A(s_A)$ and $p_B(s_B)$. Equation (2) can be written alternatively as $w_i^* = p_i w_{Bi} + (1 - p_i) w_{Ai}$, with the obvious definition of the weights w_{Bi} and w_{Ai} associated with the samples s_B and s_A . Thus, the class of weighting schemes defined by equation (2) consists essentially of different weighted combinations of the weights in the original samples s_B and s_A . The limits on the values of p_i ensure that the weight w_i^* will be nonnegative. Note that the intractable weight $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$, for $i \in s \cap B$, used in (1) is a special case of w_i^* with $p_i = \pi_{Bi}(\pi_{Ai} + \pi_{Bi})^{-1}$.

Evidently, the weighting scheme defined by (2) does not eliminate duplicate units that fall in both samples. If the operational constraint to exclude from s_A individuals already selected in s_B is imposed, the second term in the right-hand side of (2) should be modified to $(1 - p_i) [\pi_{Ai}(1 - \pi_{Bi})]^{-1} I\{i \in s_{ab}, i \notin s_B\}$ to ensure that $E(w_i^*) = 1$. This, however, may be impossible to do since it requires that the inclusion probabilities of the sampled units be known over both frames. Note also that under the constraint of excluding duplicate units, the two samples will not be independent. Nevertheless, as it is assumed that both probabilities π_{Ai} and π_{Bi} are small, the probability of duplicate units will be negligibly small, and hence any bias resulting

from using the tractable weighting scheme defined by (2) would also be negligible. On this assumption, the two indicator variables in (2) should be understood to satisfy $I\{i \in s_B\} I\{i \in s_{ab}\} = 0$.

The question arises now as to an optimal choice of p_i , for any $i \in s \cap B$, according to some criterion of optimal weighting for the combined sample. One approach is to choose the p_i to minimize the variance of the estimated total $\hat{Y}_A = \sum_B w_i^* y_i + \sum_a w_i y_i$, where $w_i = (\pi_{Ai})^{-1} I\{i \in s_a\}$ for $i \in a$. However, minimization of the variance of \hat{Y}_A with respect to p_i for all $i \in s \cap B$ is not tractable. A simpler option is to restrict the class of weighting schemes defined by equation (2) to one in which the weight adjustment factors are specified not at the unit level but rather at a higher level, which may be a stratum or the entire overlap frame B . Further discussion on the level of adjustment is deferred to the last part of this subsection. It suffices for the development of the weighting procedure to consider next the case involving a uniform weight adjustment factor p for the entire frame B .

Determination of the value of p . Issues of practicality and efficiency.

The class of weighting schemes defined by equation (2) for the frame B , with uniform weight adjustment factor p , generates a class of unbiased estimators for the overall total Y_A of the form

$$\hat{Y}_A^p = p \hat{Y}_{s_B} + (1 - p) \hat{Y}_{s_{ab}} + \hat{Y}_{s_a}, \quad (3)$$

where \hat{Y}_{s_B} and $\hat{Y}_{s_{ab}}$ are independent Horvitz-Thompson estimators of Y_B based on s_B and s_{ab} , respectively, and \hat{Y}_{s_a} is the Horvitz-Thompson estimator of Y_a based on s_a . The limit values of p yield two special cases of the estimator \hat{Y}_A^p , in both of which the overlap domain total Y_B is estimated from one panel only. When p is set equal to zero in (3), the resultant trivial estimator \hat{Y}_A^p for the entire population is based only on s_A . More notable is the case with p set equal to one in (3). The implied simple unbiased estimator $\hat{Y}_A = \hat{Y}_{s_B} + \hat{Y}_{s_a}$ would be the natural estimator in a panel survey with one panel and a supplementary cross-sectional sample, with the units in that sample being "screened" and only the units in the domain of new entrants being enumerated. In such a context this simple estimator would be a special case of a "screening" multiple frame estimator, the special feature being the temporal nature of the non-overlap frame domain a . In the present context the screening estimator appears inefficient because information in the sample domain s_{ab} is not utilized. Better use can be made of data from both panels by combining s_B and s_{ab} , using an optimal p that is based on the minimization of the variance of \hat{Y}_A^p . The optimal value of p is given by

$$p = \frac{\text{Var}(\hat{Y}_{s_{ab}}) + \text{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a})}{\text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_{ab}})}. \quad (4)$$

The variance and covariance terms in (4) are unknown, but could be estimated from the sample data, in which case the chosen p would actually minimize the estimated variance of \hat{Y}_A^p . There are many drawbacks associated with this choice of value for p . Generally, estimation of the optimal p is not easy; in surveys with more than two panels it would be very inconvenient to estimate the required set of such weight adjustments. Also, a sample estimate of the optimal p in (4) adds variability to the estimator \hat{Y}_A^p , and complicates the estimation of its variance. Moreover, the dependency of the estimated optimal p on the sample data entails $E(w_i^*) \neq 1$ for $i \in B$, which disturbs the unbiasedness of the estimator (3). It is to be noted that the condition $E(w_i^*) = 1$ is also necessary for the validity of the weight share method (see section 4) to hold when applied to the combined sample s at any wave after the selection of the second panel.

An alternative choice for the value of p is based on the minimization of the variance of the common-frame component $\hat{Y}_B^p = p \hat{Y}_{s_B} + (1-p) \hat{Y}_{s_{ab}}$ of the estimator \hat{Y}_A^p in (3). This restricted minimization, which ignores the typically small domain estimator \hat{Y}_{s_a} , gives the value

$$p' = \frac{\text{Var}(\hat{Y}_{s_{ab}})}{\text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_{ab}})}, \quad (5)$$

which is independent of the covariance term, and always lies between zero and one. Minimizing the variance of \hat{Y}_B^p conditional on the realized value of the random size n_{ab} of the sample domain s_{ab} , then using the well-known variance formula for the estimator of a total under simple random sampling, and disregarding finite population corrections, it can be shown that (5) may be approximated by

$$\hat{p}' = \frac{n_B/d_B}{n_B/d_B + n_{ab}/d_{ab}}, \quad (6)$$

where n_B is the size of the sample s_B , and d_B , d_{ab} are the design effects associated with s_B and s_{ab} . The calculation of the value of \hat{p}' requires estimates of the two design effects, which need not be based on s_B and s_{ab} . Suitable approximate values of d_B and d_{ab} may be available from other surveys with the same sampling designs as the two panels. However, because of the dependency of \hat{p}' on the characteristic y through d_B and d_{ab} , a different set of weights needs to be calculated for each characteristic of interest. Besides making the estimation process operationally inconvenient, the different sets of weights may lead to inconsistencies among estimates. A compromise solution is to obtain approximate values of d_B and d_{ab} preferably for a count variable associated with a large population domain. A similar compromise solution is implicit in the approach of Skinner and Rao (1996) to estimation in dual frames. It is to be noted that since \hat{p}' depends on the characteristic y only through the ratio d_B/d_{ab} , the loss of efficiency for

estimators of totals of other characteristics should not be substantial. It is to be noted further that because of the time lag between the selection of the two panels, the design effects will be different, and thus present in (6), even when the sampling designs for the two panels are identical. By using estimates of the design effects from external sources the randomness of \hat{p}' is due only to the random size of the sample domain s_{ab} . Since the size of the sample s_A is usually very large, and the size of the overlap frame B is typically only a little smaller than the size of the complete frame A , the size n_{ab} of the sample domain s_{ab} must be nearly constant, and thus the unbiasedness condition $E(w_i^*) = 1$ will hold approximately.

Some loss of efficiency will be incurred by ignoring \hat{Y}_{s_a} in deriving an optimal value for p , but this loss may be insignificant given the relatively very small size of the domain a in most household panel surveys, because of the typically small time lag between panels. To assess this loss of efficiency, let $\hat{Y}_A^{p'}$ and \hat{Y}_A^p denote the estimator \hat{Y}_A^p in (3) with the value of p given by (4) and (5), respectively. Then, a simple calculation gives

$$\begin{aligned} \text{Var}(\hat{Y}_A^{p'}) - \text{Var}(\hat{Y}_A^p) &= \frac{\text{Cov}^2(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a})}{\text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_{ab}})} \\ &\leq \frac{\text{Var}(\hat{Y}_{s_{ab}})\text{Var}(\hat{Y}_{s_a})}{\text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_{ab}})} \\ &= p' \text{Var}(\hat{Y}_{s_a}), \end{aligned}$$

so that an upper bound for the efficiency loss can be obtained as

$$\frac{\text{Var}(\hat{Y}_A^{p'}) - \text{Var}(\hat{Y}_A^p)}{\text{Var}(\hat{Y}_A^p)} \leq p' \frac{\text{Var}(\hat{Y}_{s_a})}{\text{Var}(\hat{Y}_A^p)}.$$

Given the usually very small size of \hat{Y}_{s_a} relative to \hat{Y}_A^p (the size of the domain a is approximately one fortieth of the size of the complete frame A in the case of the SLID) it appears that the loss of efficiency will be very small in most panel household surveys.

An interesting question is whether or not $\hat{Y}_A^{p'}$ is more efficient than the simple "screening" estimator $\hat{Y}_A = \hat{Y}_{s_B} + \hat{Y}_{s_a}$, whose variance is $\text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_a})$. It can be readily shown that $\text{Var}(\hat{Y}_A^{p'}) < \text{Var}(\hat{Y}_{s_B}) + \text{Var}(\hat{Y}_{s_a})$ if $2\text{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) < \text{Var}(\hat{Y}_{s_B})$. This condition certainly holds if the covariance of $\hat{Y}_{s_{ab}}$ and \hat{Y}_{s_a} is negative, which may be the case if the estimated characteristic differs between immigrants versus non immigrants. In general, this covariance may actually be positive because $\hat{Y}_{s_{ab}}$ and \hat{Y}_{s_a} are based on the same sampled area clusters. In that case too, however, the condition will most likely hold, given the magnitude of $\text{Var}(\hat{Y}_{s_B})$ relative to $\text{Var}(\hat{Y}_{s_{ab}})$, and the magnitude of $\text{Var}(\hat{Y}_{s_{ab}})$ relative to $\text{Var}(\hat{Y}_{s_a})$. Indeed, the

sizes of the panel samples s_B and s_A are typically equal by design, although the effective panel sizes (i.e., realized sizes at any wave, adjusted for design effects) may be considerably different due to different attrition rates and design effects for the two panels. Also, with the sizes of the sample domains s_{ab} and s_a roughly proportional to the corresponding population domain sizes, $\text{Var}(\hat{Y}_{s_a})$ will be many times, say k , smaller than $\text{Var}(\hat{Y}_{s_{ab}})$. Then,

$$\begin{aligned} 2 \text{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) &\leq 2 \sqrt{\text{Var}(\hat{Y}_{s_{ab}}) \text{Var}(\hat{Y}_{s_a})} \\ &= 2 \frac{\text{Var}(\hat{Y}_{s_{ab}})}{\sqrt{k}}, \end{aligned}$$

so that a sufficient condition for the estimator $\hat{Y}_A^{p'}$ to be more efficient than the "screening" estimator is

$$2 \frac{\text{Var}(\hat{Y}_{s_{ab}})}{\sqrt{k}} < \text{Var}(\hat{Y}_{s_B}).$$

The interpretation of this is that the sample domain s_{ab} is not to be ignored when estimating Y_A if $\text{Var}(\hat{Y}_{s_B})$ is not too small relative to $\text{Var}(\hat{Y}_{s_{ab}})$. The condition is ordinarily satisfied in panel household surveys. An additional argument in favour of including s_{ab} in estimation is its better quality relative to s_B , since the latter is more liable to the potential bias effect of sample attrition.

The simple approximate weight adjustment factor \hat{p}' given by expression (6) affords an efficient combination of panel samples, accounting for the precision of $\hat{Y}_{s_B}^{p'}$ relative to that of \hat{Y}_{s_a} through the effective sample sizes n_B^*/d_B and n_{ab}^*/d_{ab} . These effective sample sizes are time-dependent, though their ratio (and hence \hat{p}') should be quite stable over the period of panel overlap. Regarding variance calculations, since n_{ab} is typically nearly non-random, the adjustment factor \hat{p}' can be conveniently treated as constant in any variance estimation procedure.

It is important to emphasize here that additional gains in efficiency will result from the incorporation of auxiliary information into the weights through a calibration weight adjustment to known population totals.

Finally, it should be remarked that if the criterion in the choice of the value of p is the minimization of the mean square error of the common-frame component $\hat{Y}_B^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_{ab}}$ of the estimator \hat{Y}_A^p , then it can be easily shown that when the biases of \hat{Y}_{s_B} and $\hat{Y}_{s_{ab}}$ are equal the optimal value of p is the same as the one given by (5). The biases are not expected to be equal, though; for instance, the different sample attrition rates for the two panels may result in different levels of bias. It is clear that the bias of the linear combination $\hat{Y}_B^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_{ab}}$, though not minimized if p is as in (5), is nevertheless smaller than the larger of the two component biases. Other complexities aside, the unavailability of good estimates for the two biases renders the criterion of minimum mean square error impracticable.

Generalization to multiple panels and discussion of alternative approaches.

The weighting procedure described above applies to the simple situation of a two-panel survey at the start of the second panel. At later survey waves an additional non-overlap frame domain, denoted by b , may be formed by returning leavers of the frame B . Units from b originally selected in the first panel were not present when the second panel was selected. Clearly, the weights in the non-overlap sample domain s_b are not to be adjusted for the purpose of combining the two panels. Furthermore, the value for p will not be affected, as it is based only on the overlap domain of the combined sample. As with ignoring the sample domain s_a in determining the value of p , ignoring the much smaller, possibly void, sample domain s_b will have negligible impact on the efficiency of derived estimators.

The simplicity of the proposed weighting procedure for the combination of two panels makes its generalization to surveys with more than two overlapping panels straightforward. The most likely generalization in practice would involve three panels. The construction of a combined cross-sectional sample would then involve the adjustment of the sampling weights of units from temporal domains of the different panels that represent a common temporal domain of the cross-sectional population. For each common temporal population domain the weight adjustment factors will be based on the relative effective sample sizes of the corresponding panel domains, in analogy with expression (6), and will add up to one. The number of common temporal frame domains, and hence the number of the corresponding independent sets of adjustment factors, will be quite small because of the high degree of nesting in the sequence of panel frames. For instance, for a three-panel survey there will be one set of three adjustment factors and one set of two.

Returning now to an earlier point, varied weight adjustment factors may be specified at a lower level of sample grouping, such as a certain stratification level. For reasons of feasibility (identical stratification for the two panels is required for that level) and operational convenience, a high level of stratification should be chosen. The natural choice is a superstratum level, at which all other weighting and estimation procedures are carried out independently for each superstratum. In the SLID, such superstrata are the Canadian provinces. The advantage of specifying weight adjustment factors at the superstratum level is improved efficiency, since an optimal or nearly optimal weight adjustment factor p can be determined for each superstratum. This will be particularly advantageous if the ratios of the effective sample sizes of the panels are very different among the superstrata, as is the case in the SLID.

Alternative estimation techniques from the general theory of multiple frame surveys with complex designs (for an account, see Skinner and Rao 1996, and Singh and Wu 1996) would produce estimators similar in form to the

estimator (3) if adapted to a multiple panel survey with overlapping panels. Such techniques, though, are not preferable in general for reasons similar to those stated in the discussion following equation (4); the “pseudo-likelihood” method of Skinner and Rao (1996) is also not applicable in surveys with more than two panels. Furthermore, while the weight adjustment proposed in this section essentially combines the panels, on the basis of an efficient combination of Horvitz-Thompson estimators, the standard multiple frame methods ordinarily combine ratio-adjusted or, more generally, calibrated estimators derived separately using the sample from each frame. In the context of a household panel survey, the components from each panel would be calibrated estimators incorporating all the weight adjustments, including the “weight share” adjustment, carried out separately for each panel. This would be in conflict with the application of the “weight share” adjustment to the combined sample, to be proposed in section 4. It is interesting to note that apart from this complication there are many possible limitations that could render a separate calibration of each panel problematic or unfeasible. It may be remarked first that a proper separate calibration of the panels is possible only when the various temporal sample domains are identifiable. Furthermore, a calibration involving the same auxiliary variables for each temporal domain of each panel would be required in order for the final weights to satisfy all calibration constraints. But since all temporal frame domains (except the one that is common to all panels) are typically very small, a calibration involving a large number of auxiliary totals (as is customary in household surveys) would not be sensible for reasons of potential bias and loss of efficiency of derived estimators. Moreover, auxiliary totals for frames of old panels that account for the loss of population units may not be available. It should also be pointed out that accurate auxiliary totals most likely would be unavailable if the frame of each panel were augmented with new entrants who live with individuals of the original frame of the panel. Such would be the situation if the “weight share” procedure, which assigns a basic weight to new entrants living with selected individuals, were to precede the combination of the panels.

Notwithstanding other difficulties, it is possible in principle to use standard multiple frame methods to combine the panels, avoiding a separate calibrating weight adjustment, with the exception of the dual-frame pseudo-likelihood method of Skinner and Rao which in the setting of Figure 1 would require a simple ratio weight adjustment for s_B , s_{ab} and s_a .

Lastly, a known drawback of various multiple frame estimators is that their optimality depends on the estimated characteristic of interest. For the proposed method this dependency appears to be weaker, because the optimal \hat{p}' in (6) depends on the particular characteristic only through a ratio of panel design effects, estimated from an extraneous source.

3.2 Non-identifiable Temporal Sample Domains

It has been assumed thus far that the units of the non-overlap sample domain $s_a (\subset s_A)$ can be identified. However, the information needed to determine whether a unit in s_A belongs to the frame domain a , of new entrants into the population after the start of the previous panel, may not be available for all units of s_A . In that situation the weighting process described above would combine the two samples s_B and s_A without distinguishing between the domains s_{ab} and s_a of s_A , so that the weights of units in s_a would also be multiplied by $1 - p$. The estimator \hat{Y}_A^p in (3) would collapse then to

$$\hat{Y}_A^p = p \hat{Y}_{s_B} + (1 - p) \hat{Y}_{s_A}. \quad (7)$$

The effect of this error is the underestimation of the total Y_a for the population domain a by the factor p . Part of the domain a , though, consists of newborns, which can be identified in s_A with certainty. Their weights could very well be excluded from the adjustment by the factor $1 - p$, but that would have no effect on cross-sectional estimation, unless newborns were part of the population of interest. Besides, adjusting the weights of newborns in s_a by the factor $1 - p$ has the desirable effect of producing a common household weight. A calibration of the weights of the combined sample to known population totals of the complete frame A will lessen the under-representation of the rest of the domain a , which consists mainly of immigrants, but some bias may still result if the survey characteristics of the members of this part of the population are quite different from those of the members of the population domain B . Unless the time lag between the selection of the two panels is quite large, the size of this part of the population is very small, relative to the total population, and the potential bias effect on overall estimates of totals should be negligible.

The optimal (*i.e.*, variance minimizing) value of p in (7) is given now by

$$p'' = \frac{\text{Var}(\hat{Y}_{s_A})}{\text{Var}(\hat{Y}_{s_A}) + \text{Var}(\hat{Y}_{s_B})}. \quad (8)$$

Disregarding finite population corrections it can be shown that (8) can be expressed as

$$\begin{aligned} \hat{p}_c'' &= \frac{n_B d_A N_A^2 S_A^2}{n_B d_A N_A^2 S_A^2 + n_A d_B N_B^2 S_B^2} \\ &= \frac{n_B d_A}{n_B d_A + c n_A d_B}, \end{aligned} \quad (9)$$

with $c = (N_B^2 S_B^2)(N_A^2 S_A^2)^{-1}$, and where n_B, n_A are the sizes of the samples s_B and s_A ; d_B, d_A are the design effects associated with s_B and s_A and the characteristic y ; N_A, N_B are the sizes of the frames A and B ; S_A^2, S_B^2 are the variances of the characteristic y in A and B . Noting that N_B may be only a little smaller than N_A (depending on the time lag between the two panels), and assuming that the unknown variances S_A^2 and S_B^2 are nearly equal, a good practical approximation of the optimal p can be obtained by simply setting c equal to one in (9). The assumption that the variances S_A^2 and S_B^2 are nearly equal is reasonable considering the magnitude of N_B relative to that of N_A . Approximate values of d_B and d_A available from other surveys with the same designs as the two panels could be used, preferably for a characteristic such as the size of a large population domain. Now, if \hat{Y}_c and \hat{Y}_1 denote the estimator \hat{Y}_A^p in (7) when the weight adjustment \hat{p}_1'' in (9) is used with the true value of c and the approximate value $c = 1$, respectively, then ignoring finite population corrections the loss of efficiency of \hat{Y}_1 relative to \hat{Y}_c can be readily shown to be

$$\frac{\text{Var}(\hat{Y}_c) - \text{Var}(\hat{Y}_1)}{\text{Var}(\hat{Y}_c)} = -\frac{(c-1)^2}{c} \hat{p}_1'' (1 - \hat{p}_1'').$$

With a value of c most likely in the neighbourhood of 1.0, the loss of efficiency will be negligible.

It is interesting to examine the efficiency of the estimator given by (7), with p'' as in (8), relative to the optimal estimator given by (3), with p as in (4), used when the domain s_a is identifiable. Let \hat{Y}_A'' and \hat{Y}_A denote these estimators, respectively. Then, using the inequality $\text{Cov}^2(\hat{Y}_{s_a}, \hat{Y}_{s_{ab}}) \leq \text{Var}(\hat{Y}_{s_a}) \text{Var}(\hat{Y}_{s_{ab}})$ it can be shown that $\text{Var}(\hat{Y}_A) - \text{Var}(\hat{Y}_A'') \geq (p'' - p') \text{Var}(\hat{Y}_{s_a})$, where p' is as in (5). As already mentioned, in general $\text{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) > 0$, so that $p'' > p'$ and hence $\text{Var}(\hat{Y}_A) \geq \text{Var}(\hat{Y}_A'')$. Therefore, notwithstanding the use of the exact values of p'' and p' in the comparison, the approach taken in this subsection may in most cases result in reduction of the variance of derived estimators. A lower bound for the gain in efficiency relative to \hat{Y}_A would then be given by

$$\frac{\text{Var}(\hat{Y}_A) - \text{Var}(\hat{Y}_A'')}{\text{Var}(\hat{Y}_A)} \geq \frac{(p'' - p')}{1 - p'}.$$

An extension of the weight adjustment procedure described above to surveys involving more than two panels with non-identifiable temporal sample domains is straightforward. There will be then as many weight adjustment factors, adding up to one, as there are panels. This very practical procedure will produce good cross-sectional estimates in multiple panel surveys in which the time lag between the selection of the panels is not large. Otherwise, the potential for bias due to the domain identification error may be of concern, mainly for estimates related to

subpopulations composed in substantial proportion of new entrants.

4. THE WEIGHT SHARE METHOD FOR THE COMBINED PANELS

This section describes the application of a weight adjustment method, known as the weight share method, to the combined panel sample at any wave after the start of the most recent panel. This weight adjustment is necessary because of the changes in the household membership after the selection of the panels.

The weight share method is a cross-sectional weighting procedure that assigns a basic weight to every individual in a panel household at any wave after the first. In particular, the weight share method, as applied to a single panel, assigns a positive weight to non-selected individuals who join households containing at least one individual selected for the original sample. Following Lavallée (1995), in this paper such households are termed longitudinal households, while the non-selected individuals living in longitudinal households are termed cohabitants. The cohabitants are distinguished into originally present cohabitants if they belong to the original (sampled) population, and originally absent cohabitants if they are new entrants to the population. Other problematic situations that can be handled by the weight share method involve non-selected households formed after the first wave by members of separate originally selected households, as well as originally selected individuals who have subsequently moved to other longitudinal households. For a detailed discussion of the weight share method for a single panel, see Kalton and Brick (1995), and Lavallée (1995). For the purpose of applying the weight share method to a multiple panel survey the following need to be considered. In multiple panel surveys, the original population for the combined panels is the union of the populations covered by the different panels at the time of their selection. Accordingly, the original sample consists of all selected units in the combined panel sample. Thus, an originally present cohabitant is an individual that was eligible for selection in any of the panels. In this approach then, at any wave after the selection of the most recent panel a cohabitant is distinguished into originally present or originally absent with respect to the original combined panel sample, not with respect to each original panel. Notably, at the first wave of a new panel, or when a top-up sample is used, all cohabitants are originally present. On the other hand, application of the weight share method separately to each panel (before combination) would require more precise information on the eligibility of the cohabitants for selection in each of the various panels, in order to distinguish the originally present cohabitants from the originally absent cohabitants and to identify the temporal domain that includes each of the cohabitants. Such information most likely would be unavailable. Moreover, combining the panels after the weight share

procedure would require a very complicated set of specifications in order to ensure that a suitable weight adjustment factor would be applied to each sampled unit. For instance, with the inclusion of the originally absent cohabitants into the panels through the weight share procedure, the frames of the panels will be different at each survey wave, thereby complicating the determination of the various temporal domains. Lastly, it should also be pointed out that in multiple panel surveys sampled individuals may move from one panel to another panel between waves during the time period of panel overlap, and non-sampled households may be formed by members of originally selected households from different panels. Thus, the panels are truly distinct (and independent) only with respect to the time of their selection.

It follows from the foregoing considerations that the weight share method for multiple panels is to be applied to the combined panel sample, and not to each panel separately. Then, with the prescribed distinction of the two types of cohabitants, the case of the weight share method for a multiple panel survey reduces to the case of a single panel survey. As a desirable consequence, the application of the weight share method to the combined sample will yield always a common weight for all members of the same household. The following is an exemplification of the proposed weight share procedure for multiple panel surveys, involving the simple case of two panels.

Starting with a survey setting as depicted in Figure 1, with two overlapping panels at the time point of the start of the second panel, let there be N individuals in the population at a later wave (time t), with N_i individuals in household \mathcal{H}_i , say; $i = 1, \dots, H$ and $\sum N_i = N$. Let M_i denote the number of individuals in household \mathcal{H}_i at time t that belong to the original population, with M_{Bi} and M_{ai} individuals from the original frame domain B and the non-overlap frame domain a , respectively, so that $M_i = M_{Bi} + M_{ai}$. Some, but not all, of the numbers M_{Bi} , M_{ai} and $N_i - M_i$ may be zero for any particular household. Now, with the random weights of individuals in B and a as defined in section 3.1, and with the weights of the $N_i - M_i$ originally absent cohabitants in \mathcal{H}_i being identically equal to zero, the weight share method defines a common weight for every individual in \mathcal{H}_i (including new members) as

$$w_i = \frac{1}{M_i} \sum_{k=1}^{M_i} w_{ik}, \quad (10)$$

where w_{ik} is the weight of the k -th household member that belongs to the original population. Clearly then $E(w_i) = 1$ for each household for which $M_i \neq 0$, whereas $E(w_i) = 0$ if $M_i = 0$, since $w_i \neq 0$ only if $M_i > 0$. For the survey characteristic y , the total for the population of individuals at time t can be expressed as $Y = \sum_{i=1}^H \sum_{k=1}^{N_i} y_{ik}$, where y_{ik} is the value of y for individual k in household \mathcal{H}_i . Then, an estimator of Y is given by

$$\begin{aligned} \hat{Y} &= \sum_{i=1}^H w_i \sum_{k=1}^{N_i} y_{ik} \\ &= \sum_{i=1}^H w_i \left[\sum_{k=1}^{M_{Bi}} y_{ik} + \sum_{k=1}^{M_{ai}} y_{ik} + \sum_{k=1}^{N_i - M_i} y_{ik} \right] \\ &= \hat{Y}_B + \hat{Y}_a + \hat{Y}_{A^c}, \end{aligned} \quad (11)$$

with w_i as in (10), with A^c denoting the set of individuals not in frame A , and with the obvious notation for the right hand side of (11). The estimator \hat{Y} in (11) is given as the sum of three estimators, \hat{Y}_B , \hat{Y}_a and \hat{Y}_{A^c} , for the totals related to the population domains B , a and A^c , respectively. The estimators \hat{Y}_B and \hat{Y}_a are unbiased, even though they are based on sets of units that may not be identical to the original samples $s_B \cup s_{ab}$ and s_a , respectively. For example, the estimator \hat{Y}_B is based on a set of units consisting of the remaining units of the original combined sample $s_B \cup s_{ab}$ from frame B , and possibly of cohabitants originally present in B . The estimator \hat{Y}_{A^c} is not unbiased for Y_{A^c} , because individuals in A^c who live in households that contain no members of the original population are not represented in the panel survey. Nevertheless, the estimator \hat{Y}_{A^c} is unbiased for the total corresponding to the rest of A^c , which is represented in the combined panels by the originally absent cohabitants. In the special case when time t coincides with the start of the second panel (or with the time of selection of a supplementary sample), $A^c = \emptyset$, $N_i = M_i$, and the estimator $\hat{Y} = \hat{Y}_B + \hat{Y}_a$ is unbiased for Y . It should be noted here that if the weights of the responding individuals at time t are adjusted for nonresponse, the relationship $E(w_i) = 1$ may hold only approximately, and in that sense the resulting estimators may be only approximately unbiased.

It is important to note that the estimator \hat{Y} in (11) can be expressed as

$$\hat{Y} = \sum_{i=1}^H w_i Y_i,$$

where $Y_i = \sum_{k=1}^{N_i} y_{ik}$ is the total for household \mathcal{H}_i . Thus, \hat{Y} is also an estimator of the household-level total at time t .

As with the weight adjustment involved in the combination of panels, the weight share adjustment may also be carried out at a superstratum level, say province, for the combined sample of each province. In this approach, those individuals who at time t reside in a province other than the one in which they resided at the time of selection of any of the panels are treated as originally absent, since they were not members of the original population of their new province. In particular, interprovincial movers (selected or non selected in their original province) who are found in longitudinal households in their new province at time t are treated as originally absent cohabitants. When a top-up sample is used at time t , these interprovincial movers are

treated as originally present cohabitants. The application of the weight share procedure separately for each superstratum enjoys certain operational and statistical advantages over the standard weight share procedure. An account of the comparative merits of the two approaches is given in Merkouris (1999).

5. INTEGRATION OF VARIOUS WEIGHT ADJUSTMENTS

In addition to the weight adjustments described so far, other adjustments to the weights of a panel household survey may also be required. The integration of the various weight adjustments is briefly outlined below.

The first adjustment, applied in relation to the original sample units, is for wave nonresponse, which arises when a sampled unit responds for some but not all of the waves for which it was eligible. For a discussion on weight adjustment for wave nonresponse, see Kalton and Brick (1995). The adjustment is made separately to the different panels at each wave.

The second adjustment is for the combination of the samples of the various panels into one sample for cross-sectional estimation. It applies to the weights of the sampled units of the panels, adjusted for wave nonresponse, and employs the method described in section 3.

The third adjustment involves the application of the weight share procedure to the combined panel sample at any wave after the start of the most recent panel, as described in section 4.

Finally, in the weight calibration adjustment the weights of the combined panel units are adjusted so as to make the estimated totals for certain auxiliary characteristics equal to known population totals for these characteristics at the current wave, which in the simple case as in Figure 1 correspond to totals of the complete frame A. In more general situations, after the selection of the most recent panel the calibration totals will include the new entrants into the population. Note that in the absence of a top-up sample the new entrants will be represented in the panels only by the originally absent cohabitants. Calibrating the weights of the combined sample to population totals of each of the different temporal domains (when the panel units from these domains can be identified) may not be feasible or sensible for reasons already noted in section 3.1.

6. SUMMARY AND CONCLUDING REMARKS

The weighting procedures described in this paper can be used to combine information from multiple panels of a repeated household survey for cross-sectional estimation in a fairly general setting involving panels with given designs; design issues regarding determination of optimal sampling fractions for the panels, in conjunction with efficient

combination of the panel data, are beyond the scope of this paper. It has been shown that although a multiple panel survey can be viewed as a special type of multiple frame survey, its distinctive dynamic character renders conventional multiple frame estimation procedures problematic or even non applicable. The proposed weighting procedures, which account for the population and panel dynamics, involve a simple weight adjustment for each panel that is proportional to the effective panel size. These procedures are operationally convenient for any number of overlapping panels, and for different situations regarding the identifiability of various temporal panel domains. Theoretical and practical issues related to the application of a weight share adjustment, to the calibration weight adjustment and to the integration of the various weighting procedures involved in a multiple panel survey have also been addressed. In particular, it has been argued that the weight adjustment for the combination of the panels should precede the weight share adjustment, with calibration being the final weight adjustment. A detailed empirical study of issues pertaining to the determination of weight adjustment factors for combining two panels of the SLID, based on the methodology of this paper, is described in Latouche *et al.* (2000). The variance of cross-sectional estimators has been discussed in this paper only in the context of efficient combination of panels. Variance estimation issues related to changes in the sample over time, particularly to moves from stratum to stratum, are discussed in Merkouris (1999). It is to be remarked, in conclusion, that the quality of a cross-sectional estimation procedure depends on the identifiability of various overlap temporal sample domains; on design features of the survey, such as the duration of (and the lag between) the panels and the use of a supplementary sample at any survey wave; and on the adequacy of the information on cohabitants required for the application of the weight share method.

ACKNOWLEDGEMENTS

The author wishes to thank Milorad Kovacevic, Michel Latouche, Pierre Lavallée and Harold Mantel for useful comments. Detailed comments and suggestions by three referees on an earlier version of this paper improved both its content and its presentation.

REFERENCES

- BANKIER, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- DEVILLE, J. C. (1998). Les enquêtes par panel : En quoi différentes des autres enquêtes? Suivi de comment attraper une population en se servant d'une autre. *Actes des Journées de méthodologie statistique*, numéro 84-85-86, 63-82.

- KALTON, G., and ANDERSON, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65-82.
- KALTON, G., and CITRO, C. F. (1993). Panel Surveys: Adding the fourth dimension. *Survey Methodology*, 19, 205-215.
- KALTON, G., and BRICK, J. M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- LATOUCHE, M., DUFOUR, J. and MERKOURIS, T. (2000). Cross-sectional weighting for the SLID: Combining two or more panels. Income Research Paper Series, 75F0002MIE6, Statistics Canada.
- LAVALLÉE, P. (1994). Ajout du second panel à l'EDTR : sélection et pondération. Internal document, Statistics Canada.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LAVIGNE, M., and MICHAUD, S. (1998). General aspects of the Survey of Labour and Income Dynamics. Working Paper SLID 98-05 E, Statistics Canada.
- MERKOURIS, T. (1999). On the weight share method for panel household surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 255-260.
- SINGH, A.C., and WU, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 69-77.
- SKINNER, C.J., and RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Producing Small Area Estimates From National Surveys: Methods for Minimizing use of Indirect Estimators

DAVID A. MARKER¹

ABSTRACT

National surveys are usually designed to produce estimates for the country as a whole and for major geographical regions. There is, however, a growing demand for small area estimates on the same attributes measured in these surveys. For example, many countries in transition are moving away from centralized decision-making, and western countries like the United States are devolving programs such as welfare from Federal to state responsibilities. Direct estimates for small areas from national surveys are frequently too unstable to be useful, resulting in the desire to find ways to improve estimates for small areas. While it is always possible to produce indirect, model-dependent, estimates for small areas, it is desirable to produce direct estimators where possible. Through stratification and oversampling, it is possible to increase the number of small areas for which accurate direct estimation is possible. When estimates are required for other small areas, it is possible to use forms of dual-frame estimation to combine the national survey with supplements in specific areas to produce direct estimates. This article reviews the methods that may be used to produce direct estimates for small areas.

KEY WORDS: Small area estimation; Direct estimation; Stratification; Oversampling; Dual-frame estimation.

1. INTRODUCTION

Throughout the world there is an increased demand for small area estimates. During the 1990s countries in transition moved away from centralized decision-making, requiring accurate estimates of local economic and demographic conditions. In the United States the Federal government has been moving responsibility for many social programs to the 50 states. Evaluating the success of such efforts requires accurate estimates for each state. Some programs such as the Small Area Income and Poverty Estimates (Citro and Kalton 2000) are required at much smaller levels of geography, for example for thousands of school districts. Regardless of the best plans of survey designers, "The client will always require more than is specified at the design stage" (Fuller 1999, page 344).

Ideally such estimates would be produced from direct (design-based) estimators. Unfortunately, at small levels of aggregation, the direct estimates are too unstable to be published and/or used for policy purposes. As a result there has been a great deal of interest in developing a range of indirect estimation techniques (Marker 1999; Rao 1999; Ghosh and Rao 1994).

This paper approaches this problem from a different perspective, how to minimize model-reliance through good survey design. It will never be possible to anticipate all survey uses, or to allocate sufficient sample sizes to all domains of interest, so indirect estimators will always be needed. It is possible, however, to make design choices that will greatly improve the ability of national surveys to support direct estimation for many small areas. Such choices can also improve the ability of surveys to be used to produce indirect estimates where they are needed. This

paper is an update of the excellent paper by Singh, Gambino and Mantel (1994) on the same topic. Design issues that will be considered include stratification and oversampling, combining multiple years of data, harmonization across surveys, dual-frame estimation, and measuring the accuracy of estimates.

2. STRATIFICATION AND OVERSAMPLING

Deciding on the optimal stratification and oversampling scheme for any national survey is a compromise across many variables of interest. Optimizing stratification and oversampling between national estimates and small area estimates should also be a compromise. By giving up some national accuracy it is often possible to greatly improve the accuracy for many small areas. Some of these small areas may then be able to support accurate design-based estimates. Other small areas will still require model assistance, but the stratification may allow for unbiased (but variable) estimates that can be incorporated into the model-based estimates. As the following example demonstrates, stratification alone is helpful, but limited, in its ability to improve small area estimates.

The United States Current Population Survey (CPS), conducted by the U.S. Census Bureau, has stratified by state and unemployment rate since 1985. However, another large Census Bureau survey, the United States National Health Interview Survey (NHIS), stratified by region, metropolitan area status, labor force data, income, and racial composition until 1994. The resulting sample sizes for individual states varied from year to year and did not support unbiased state-level estimates. Due to random sampling, from 1985

¹ David A. Marker, Westat, 1650 Research Blvd., Maryland, U.S.A. 20850, e-mail: DavidMarker@Westat.com.

to 1994 two states did not have any sample included in the NHIS. This would not have happened with state stratification.

Beginning in 1995 the NHIS stratification scheme was replaced by state and metropolitan status. Table 1 summarizes the number of states that have sufficient sample size in the 1995 NHIS to achieve various levels of accuracy for four different key health measures. The NHIS completes interviews with approximately 44,000 households containing 100,000 individuals. With a very strict constraint of a 10 percent coefficient of variation (CV) less than 10 states meet the standard for three of the four variables. Over half of the states meet the more lenient 30 percent CV for all four variables, but even this standard is not met for all states.

Figure 1 presents the ability of the NHIS to meet these accuracy standards for generic questions with prevalence levels of 0.01, 0.05, 0.10, 0.15, and 0.20 and design effects ranging from 1.00 to 6.00. (This variation in design effects is found on the NHIS, depending on the intra-household correlation and other clustering.) For prevalence rates above 10 percent, almost all states can achieve the 30 percent criterion even for the largest design effects. However, there is a significant drop off in the number of states as the criterion is tightened, the design effect increases, or the prevalence rate drops. For rare events with even moderate design effects less than half the states can meet the weakest criterion and hardly any can make the tightest.

Table 1
Summary of the Number of States (out of 51, Including the District of Columbia) That Have the Required 1995 NHIS Sample Size to Achieve a CV of 30-, 20-, and 10-Percent for Four Selected Variables (44,000 Households, 100,000 Individuals)

Coefficient of Variation (CV)	Percent uninsured: all ages (p = 13.5%)	Percent uninsured: under 19 (p = 12.2%)	Percent uninsured: low income children (p = 20.4%)	Percent smokers: 18 and over (p = 25.2%)
30-percent	42	31	28	45
20-percent	31	13	10	36
10-percent	7	2	2	14

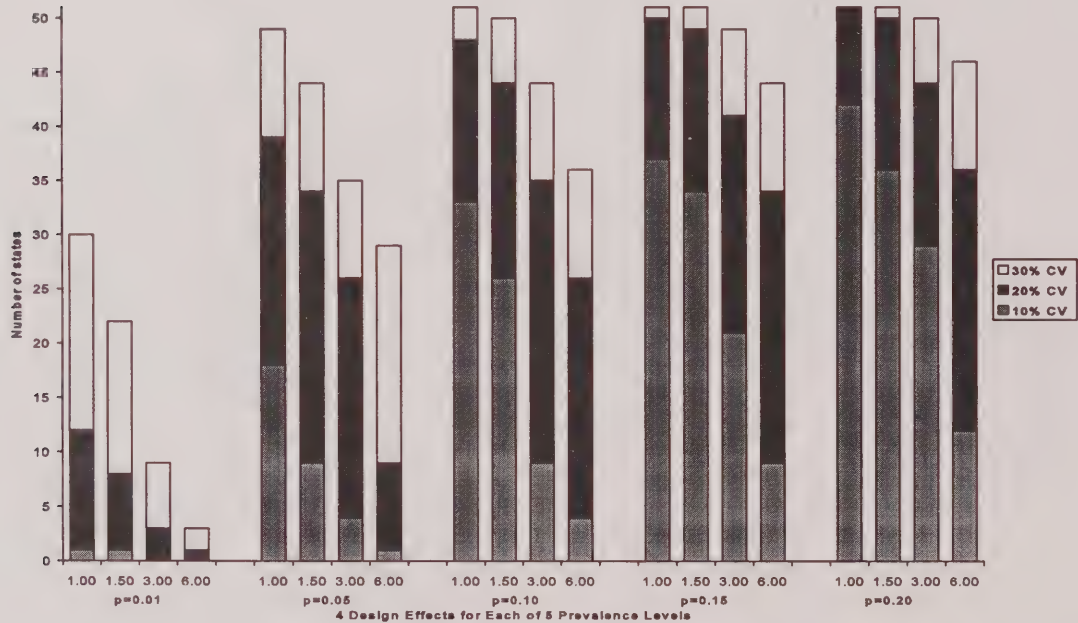


Figure 1. Number of States Meeting CV Criteria for 1995 NHIS (44,000 Households, 100,000 Individuals)

Stratification by small area assures a fixed sample size will be assigned to each small area, and thereby fixes the accuracy associated with direct estimates. Without such stratification, it may even be impossible to produce unbiased estimates for small areas that do contain some sample, because the probabilities of selection for sampled cases are a function of their entire stratum, both inside and outside the small area. For example, this can occur when part of a small area is in a stratum that crosses small area boundaries, and the sampled PSUs are in other small areas. To produce direct estimates requires either collapsing strata boundaries or small area boundaries.

By oversampling small areas it is possible to significantly improve the accuracy of direct estimates for these areas, while only incurring a minimal loss in accuracy for national estimates. As a simple example, consider a national survey with 5,000 respondents but where under a random sampling scheme 10 of the small areas would only receive 100 cases each. Alternatively one could double the sample size to 200 in each of these small areas while retaining the national sample size of 5,000. The effective sample size for national estimates would be reduced by this oversampling, but would remain more than 4,000, so the CV of national estimates would increase less than 10 percent. The CV for estimates in each of the 10 small areas would decrease 30 percent because the sample size was doubled.

Beginning in 1999 the U.S. National Household Survey on Drug Abuse has combined stratification and oversampling to produce direct estimates for every state (Chromy, Bowman and Penne 1999).

Singh *et al.* (1994) provided an example of oversampling small areas in the Canadian Labour Force Survey. Seventy percent of the sample was allocated to provide optimal national and provincial estimates. The remaining 30 percent were used to supplement small areas to improve their estimates. National CVs were increased between 10 and 20 percent by this compromise design, but unemployment insurance regions' estimates had CV reductions as large as 50 percent.

A similar design was used for the 2000 Danish Health and Morbidity Survey. The survey included two national samples, each of 6,000 respondents. An additional 8,000 respondents were distributed to assure that at least 1,000 respondents would be in each county.

The effect of oversampling on CVs can also be seen by comparing the 1996 CPS and 1995 NHIS with America's 1996 Survey of Income and Program Participation (SIPP). The CPS not only stratified by state, it also oversampled smaller states. The NHIS stratified by state but didn't oversample based on geography (minority groups were oversampled, but they tend to be located in the more populous states). In contrast, the SIPP did not stratify by state nor did it oversample. The ratio of the largest to smallest state sample size was 11:1 for CPS, 60:1 for SIPP, and 110:1 for NHIS. The corresponding ratio of CVs was

3.5:1, 7.5:1, and 10.5:1. Oversampling resulted in the CVs for the smallest states being reduced by almost a factor of two-thirds!

It is important to remember that oversampling based on geography doesn't necessarily reduce the variability in other domains of interest, for example demographic subgroups. The ratios of largest to smallest state sample sizes in the CPS were 15:1 for children, 20:1 for the elderly, 500:1 for Blacks, and 800:1 for Hispanics.

The 1994 U.S. National Employer Health Insurance Survey (NEHIS) oversampled smaller states to balance the need for accurate state and national estimates. The overall sample of 40,000 establishments had to be spread across all 51 states to provide direct estimates for all states. Three options were considered:

- Option A: Optimal national allocation (based on total employment in the state) yielded very small sizes in some states.
- Option B: Equal allocation to all states yielded inefficient national estimates.
- Option C: Minimum 400 completes per state (allocate based on number of employees to the 0.3 power).

The corresponding ratio of largest to smallest state CVs were 7.2:1 for Option A, 1:1 for Option B, and 1.8:1 for Option C. Compared to Option C, the national CV with Option A was 17 percent lower, but with Option B was 22 percent higher. Option C was selected over Option A since it reduced the variation in state CVs by a factor of 4 while only moderately increasing the national CV.

3. COMBINING MULTIPLE YEARS

An inexpensive way to increase the sample sizes in small areas is to combine cycles of a repeated survey. Combining k years of an annual survey increases the effective sample size not quite k times. The reason for this is that usually consecutive years of the same survey are conducted in the same primary sampling units (PSUs) and even adjacent area segments. This results in some correlation between years, somewhat reducing the effective sample size.

One drawback to combining multiple years is that such estimates are slow to detect changes across time. If time series are a prime interest, alternative methods must be used to increase the sample size.

Table 2 shows for the 1995 NHIS how many states can achieve different levels of accuracy by aggregating across two or three years. Aggregation clearly helps achieve CVs of 30 and 20 percent. Even aggregating 3 years can't help many states achieve a CV of 10 percent.

Table 2

Summary of the Number of States (out of 51) That Have the Required 1995 NHIS Sample Size to Achieve a CV of 30-, 20-, and 10-Percent; Aggregating Multiple Years for Four Selected Variables (44,000 Households, 100,000 Individuals).

	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
30-percent CV				
1 year	42	31	28	45
2 years	46	35	36	50
3 years	49	41	37	51
20-percent CV				
1 year	31	13	10	36
2 years	36	29	24	44
3 years	42	31	31	46
10-percent CV				
1 year	7	2	2	14
2 years	14	3	3	25
3 years	22	7	4	32

4. HARMONIZATION ACROSS SURVEYS

Harmonizing questions across surveys is another inexpensive way to improve estimation. Eurostat has been making a major effort to harmonize a number of surveys both between countries and within. The European Community Household Panel Survey (ECHP) is an attempt to collect consistent information across the member countries. Similar standardization is ongoing in each country's Labour Force Survey. This harmonization across countries improves international comparisons.

Harmonizing across surveys of the same population increases sample sizes, improving small area estimates. Statistics Finland has been harmonizing the process for collecting income and other variables in its surveys. The Permanent Survey on Living Conditions (POLS) at Statistics Netherlands uses a common procedure for collecting basic information in a series of social surveys.

Even if the questionnaire wording is consistent across surveys, the data may not be completely comparable. Different modes of data collection can cause differences, as can the placement of questions (Groves 1989).

5. DUAL-FRAME ESTIMATION

In some situations it is possible to supplement an in-person survey with telephone data collection, thereby increasing the sample size in a small area at more limited expense. The Dutch Housing Demand Survey is a national in-person survey. To produce small area estimates telephone supplementation is used in over 100 municipalities. Table 3 shows the size of the national in-person survey, telephone supplement, and total sample in ten selected municipalities.

Table 3

Dual-Frame Completes for Municipalities in the Dutch Housing Demand Survey

Municipality	In-Person National Survey	Telephone Supplement	Total
Leek	56	569	625
Marum	29	299	328
Slochteren	44	456	500
Zuidhorn	54	558	612
Emmen	770	224	994
Avereest	134	465	599
Bathmen	24	506	530
Dalfsen	157	466	623
Deventer	316	335	651
Diepenveen	47	336	383

Sirken and Marker (1993) described dual-frame estimation for the U.S. National Health Insurance Survey (NHIS) based on its 1985-94 design. Table 4 examines the same idea for the current design implemented beginning in 1995. The table compares the ability to produce state estimates with national in-person survey interviews and with unbiased dual-frame estimation using an unlimited number of supplemental telephone interviews. (Up to 100, 200, and 2,000 telephone interviews per state are required to achieve CVs of 30-, 20-, and 10- percent, respectively.) When a small area has a large percentage of households without telephones, no amount of telephone supplementation may be sufficient to achieve unbiased estimates with the desired accuracy.

In such situations, it may only be possible to achieve a desired level of accuracy using a potentially biased estimator that combines all data regardless of the mode of collection. The relative root mean square error (RRMSE) must then be used instead of the CV to measure accuracy. However, for some characteristics households with

telephones have different expectations than households without telephones. In such situations the bias can again prevent achieving the desired accuracy. The bias for each of these variables was estimated by comparing NHIS responses from households with and without telephones. Table 5 shows how the number of states for which a 10 percent RRMSE can be achieved varies by question, a function of the bias in telephone households and the telephone penetration rate in each state.

Small areas with high telephone penetration rates, for characteristics with different expectations for telephone and non-telephone households, are better able to produce accurate estimates using an unbiased dual-frame estimator. Small areas with lower penetration rates, for characteristics with similar telephone and non-telephone households, produce more accurate estimates with a potentially biased dual-frame estimator. Using the appropriate dual-frame estimator for a given small area and characteristic can allow accurate estimates to be produced for a large percentage of small areas.

Table 4
The Number of States Able to Achieve 30-, 20-, 10-Percent CV With the 1995 NHIS Area Sample Only, With Unbiased Dual-Frame Estimation Using a RDD Supplement, or not at All, for Four Specific Variables

CV	Data sources	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
30%	With area sample only	42	31	31	46
	With RDD supplement	9	20	19	5
	Unable to meet requirement	0	0	1	0
20%	With area sample only	32	15	10	37
	With RDD supplement	19	35	40	14
	Unable to meet requirement	0	1	1	0
10%	With area sample only	8	2	2	15
	With RDD supplement	40	41	39	36
	Unable to meet requirement	3	8	10	0

Table 5
The Number of States Able to Achieve 10-Percent RRMSE With the 1995 NHIS Area Sample Only, With a RDD Supplement, or not at all, for the Four Specific Variables

Data source	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
With area sample only	8	2	2	15
With RDD supplement				
Unbiased Estimator	40	41	39	36
Biased Estimator	30	47	49	35
Unable to meet requirement				
Unbiased Estimator	3	8	10	0
Biased Estimator	13	2	0	1

6. IMPROVING POINT AND VARIANCE ESTIMATION

When sufficient sample size exists to produce small area estimates there are additional steps that can be taken to improve their accuracy. SIPP does not stratify by state, to improve state estimates it reweights the estimates to control totals at the state level. This is very important when the stratification doesn't match the analytic domains. The use of control totals also improves subpopulation (*e.g.*, demographic) size estimates for the small areas. However, it is not possible to control as many subpopulations in a small area as can be done at the national level, due to the smaller sample sizes.

There are also many techniques to improve variance estimation for small areas. Typically there will be very few sampled PSUs in a given small area. This provides few degrees of freedom for estimating between-PSU (or total) variance. One solution is to average estimates of variance across small areas, but this covers up the fact that estimates are generally much better for some areas than for others. Alternatively generalized variance functions (GVFs) can be used to smooth variance estimates.

A preferable solution is to address small area variance estimation at the design stage. Increasing the number of PSUs, with a corresponding reduction in sample size in each PSU, can significantly improve both point and variance estimation, often at little extra cost. Singh *et al.* (1994) suggested increasing the number of PSUs to control sample sizes in unplanned small areas. Remembering Fuller's observation that "The client will always require more than is specified at the design stage," it is impossible to anticipate all small areas of interest. By having more PSUs the likelihood is increased that actual data will have been collected from unplanned analytic domains.

Kalton (1994) suggested a second reason for increasing the number of PSUs. His concern was that more PSUs per small area would greatly increase the stability of variance estimates. This is true even in very large national surveys with many PSUs. The NHIS was redesigned in 1995 increasing the number of PSUs from 196 to 359. Of these 359 PSUs 264 were noncertainty PSUs. This still resulted in only 7 states having more than 8 noncertainty PSUs. While direct variance estimation for individual states is still problematic for most states, there is an increased opportunity to develop average variance estimates for groups of states with common characteristics, rather than having to group all states together in a national average.

7. SUMMARY

There will always be a need for indirect small area estimation methods since the entire set of analytic domains is never known in advance. This need for small area estimates is growing around the world. There are, however, many actions that can be taken at the design stage to improve direct small area estimates, both point estimates and variance estimates. These steps include stratification consistent with known analytic domains, oversampling smaller areas, and increasing the number of PSUs. Given the data it is often possible to combine data from multiple years, from other surveys with whom questions have been harmonized, and through dual-frame estimation techniques. These steps will both reduce the need for indirect estimates and improve the accuracy of those estimates when they are required.

REFERENCES

- CHROMY, J.R., BOWMAN, K.R. and PENNE, M.A. (1999). The National Household Survey on Drug Abuse Sample Design Plan. Prepared for the Substance Abuse and Mental Health Services Administration, Rockville Maryland.
- CITRO, C.F., and KALTON, G. (2000). Small-area Estimates of School-age Children in Poverty: Evaluation of Current Methodology. National academy press, Washington, D.C.
- FULLER, W.A. (1999). Environmental surveys over time. *Journal of agricultural, Biological, and Environmental Statistics*, 4, 331-345.
- GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.
- KALTON, G. (1994). Comments on Singh, Gambino and Mantel. *Survey Methodology*, 20, 18-20.
- MARKER, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.
- RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.
- SIRKEN, M.G., and MARKER, D.A. (1993). Dual frame sample surveys based on NHIS and state RDD surveys. *Proceedings of the 1993 Public Health Conference on Records and Statistics*.

A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data

HIROSHI SAIGO, JUN SHAO and RANDY R. SITTER¹

ABSTRACT

In this paper, we discuss the application of the bootstrap with a re-imputation step to capture the imputation variance (Shao and Sitter 1996) in stratified multistage sampling. We propose a modified bootstrap that does not require rescaling so that Shao and Sitter's procedure can be applied to the case where random imputation is applied and the first-stage stratum sample sizes are very small. This provides a unified method that works irrespective of the imputation method (random or nonrandom), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). In addition, we discuss the proper Monte Carlo approximation to the bootstrap variance, when using re-imputation together with resampling methods. In this setting, more care is needed than is typical. Similar results are obtained for the method of balanced repeated replications, which is often used in surveys and can be viewed as an analytic approximation to the bootstrap. Finally, some simulation results are presented to study finite sample properties and various variance estimators for imputed data.

KEY WORDS: Hotdeck; Percentile method; Monte Carlo; Imputation; Bootstrap sample size.

1. INTRODUCTION

Item nonresponse is a common occurrence in surveys and is usually handled by imputing missing item values. The various imputation methods used in practice can be classified into two types: deterministic imputation, such as mean, ratio and regression imputation, typically using the respondents and some auxiliary data observed on all sampled elements; and random imputation. In both cases the imputation is often applied within imputation classes formed on the basis of auxiliary variables. This article focuses on random imputation.

Typically, random imputation is done in such a way that applying the usual estimation formulas to the imputed data set produces asymptotically unbiased and consistent survey estimators (*e.g.*, means, totals, quantiles). More details about random imputation are provided in section 2. It is common practice to also treat the imputed values as true values when estimating variances of survey estimators. This leads to serious underestimation of variances if the proportion of missing data is appreciable, and to poor confidence intervals.

There have been some proposals in the literature to circumvent this difficulty. For random imputation, Rubin (1978) and Rubin and Schenker (1986) proposed the multiple imputation method to account for the inflation in the variance, which can be justified from a Bayesian perspective (Rubin 1987). Adjusted jackknife methods for variance estimation have been proposed for both random and deterministic imputations (Rao and Shao 1992; Rao 1993; Rao and Sitter 1995; Sitter 1997), under stratified multistage sampling. However, it is well known that the

jackknife cannot be applied to non-smooth estimators, *e.g.*, a sample quantile or an estimated low income proportion (Mantel and Singh 1991).

There are two methods available for handling randomly imputed data for both smooth and non-smooth estimators: the adjusted balanced repeated replication (BRR) methods proposed by Shao, Chen and Chen (1998); and the bootstrap method proposed by Shao and Sitter (1996) (see also Efron 1994) with a re-imputation step to capture the imputation variance. The bootstrap method is more computer intensive but is easy to motivate and understand, and provides a unified method that works irrespective of the imputation method (random or nonrandom), the type of $\hat{\theta}$ (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation).

In this article we continue the work by Shao and Sitter (1996). First, we show in section 3 how Shao and Sitter's bootstrap procedure can be modified to handle very small stratum sizes (*e.g.*, two psu's per stratum). Second, we discuss in section 4 the proper Monte Carlo approximation to the bootstrap estimators, a problem for which more care is needed when random re-imputation is applied than is typical. This has no detrimental effect on bootstrap confidence intervals based on the percentile method, but if done incorrectly, will cause the bootstrap-*t* to perform poorly. Third, we consider a BRR variance estimation method with a re-imputation step, which can be viewed as an analytic and symmetric approximation to the bootstrap method. Finally, we present some simulation results to study properties of various bootstrap and BRR variance estimators.

¹ Hiroshi Saigo, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050 Japan; Jun Shao, Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706, USA; Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

2. STRATIFIED MULTISTAGE SAMPLING AND RANDOM IMPUTATION

Though the methods discussed in this article can be more generally applied, we restrict attention to the commonly used stratified multistage sampling design. Suppose that the population contains H strata and in stratum h , n_h clusters are selected with probabilities p_{hi} , $i = 1, \dots, n_h$. Samples are taken independently across strata. In the case of complete response on item y , let

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / (n_h p_{hi})$$

be a linear unbiased estimator of the stratum total Y_h , where \hat{Y}_{hi} is a linear unbiased estimator of the cluster total Y_{hi} for a selected cluster based on sampling at the second and subsequent stages. A linear unbiased estimator of the total, $Y = \sum Y_h$, is given by $\hat{Y} = \sum \hat{Y}_h$, which may be written as

$$\hat{Y} = \sum_{(hik) \in s} w_{hik} y_{hik}, \quad (1)$$

where s is the complete sample of elements, and w_{hik} and y_{hik} respectively denote the sampling weight and the item value attached to the (hik) -th sampled element.

Often a survey estimator, $\hat{\theta}$, can be expressed as a function of a vector of estimated totals as in (1). If one is interested in the population distribution function, it can be estimated by $\hat{F}_n(t) = \sum_s w_{hik} I(y_{hik} \leq t) / \hat{U}$, where $I(\cdot)$ is the usual indicator function and $\hat{U} = \sum_s w_{hik}$. Some non-smooth estimators that are of interest are the p -th sample quantile, $\hat{F}^{-1}(p)$, where \hat{F}^{-1} is the quantile function of \hat{F} , and the sample low income proportion $\hat{F}[(1/2)\hat{F}^{-1}(1/2)]$.

Suppose that the value y_{hik} is observed for $(hik) \in s_r \subset s$, termed a respondent, while for others, $(hik) \in s_m$, it is missing, termed a nonrespondent, with $s = s_r \cup s_m$. When there are missing data, it is common practice to use $\{y_{hik} : (hik) \in s_r\}$ to obtain imputed values \tilde{y}_{hik} for $(hik) \in s_m$ and then treat these imputed values as if they were true observations and estimate Y with

$$\hat{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} \tilde{y}_{hik}. \quad (2)$$

In practice, the accuracy of the imputation is improved by first forming several imputation classes using control variables observed on the entire sample, and then imputing within imputation class. For simplicity we consider a single imputation class.

Random imputation entails imputing the missing data by a random sample from the respondents, or, in the presence of auxiliary data, by using a random sample of residuals. If the imputation is suitably done, the estimator \hat{Y}_I in (2) is asymptotically unbiased and consistent, although it is not as efficient as \hat{Y} in (1). Throughout this article, we assume that, either

within each imputation cell, the response probability for a given variable is a constant, the response statuses

for different units are independent, and imputation is carried out within each imputation cell and independently across the imputation cells,

or

within each imputation cell, the response probability of a given variable does not depend on the variable itself (but may depend on the covariates used for imputation), imputation is carried out independently across the imputation cells, and within an imputation cell, imputation is performed according to a model that relates the variable being imputed to the covariates used for imputation.

We also assume the same asymptotic setting as that in Shao *et al.* (1998). Thus, consistency (or asymptotic unbiasedness) refers to convergence of estimators (or expectations of estimators) under the assumption in Shao, *et al.* (1998), as the first-stage sample size $n = \sum n_h$ increases to infinity.

There are many methods of random imputation. We consider only two in this article: the weighted hotdeck considered in Rao and Shao (1992), which we refer to simply as random imputation, and the adjusted weighted hotdeck proposed in Chen, Rao and Sitter (2000), which we refer to as adjusted random imputation. Our results can be easily extended to random imputation with residuals in the presence of auxiliary data (e.g., random regression imputation). Generalizations to other types of random imputation may be possible, but will not be considered here.

Random imputation randomly selects donors, \tilde{y}_{hik} from $\{y_{hik} : (hik) \in s_r\}$ with replacement with probabilities w_{hik} / \hat{T} , where $\hat{T} = \sum_s w_{hik}$. In this case $E_I(\hat{Y}_I) = (\hat{S} / \hat{T}) \hat{U} = \hat{Y}_r$, a ratio estimator which is asymptotically unbiased and consistent for Y , where $\hat{S} = \sum_{s_r} w_{hik} y_{hik}$. Here E_I denotes expectation under the random imputation. The variance of \hat{Y}_I is larger than the variance of \hat{Y}_r because of the random imputation. However, the distribution of item values in the imputed data set is preserved.

Adjusted random imputation simply uses $\tilde{\eta}_{hik} = \tilde{y}_{hik} + (\hat{S} / \hat{T} - \tilde{S} / \tilde{T})$ as the imputed values instead of \tilde{y}_{hik} , where $\tilde{S} = \sum_{s_m} w_{hik} \tilde{y}_{hik}$, $\tilde{T} = \sum_{s_m} w_{hik}$ and \tilde{y}_{hik} are the imputed values from random imputation. Chen *et al.* (2000) show that this method completely eliminates the variability due to the random imputation for estimating the population total. That is $\tilde{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} \tilde{\eta}_{hik} = \hat{Y}_r$. The method also retains the distribution of item values in the imputed data set. However, the resulting imputed values need not be actual realizations.

An imputed estimator of the distribution function under random imputation is given by

$$\hat{F}_I(t) = \left[\sum_{s_r} w_{hik} I(y_{hik} \leq t) + \sum_{s_m} w_{hik} I(\tilde{y}_{hik} \leq t) \right] / \hat{U}. \quad (3)$$

An imputed estimator of the distribution function under adjusted random imputation, denoted $\tilde{F}_I(t)$, is simply obtained by replacing \tilde{y}_{hik} in (3) by $\tilde{\eta}_{hik}$. For estimating the

distribution function, adjusted random imputation does not eliminate the imputation variance as it does for estimating the total. However, Chen *et al.* (2000) show that it does significantly reduce the imputation variance when compared to random imputation. Both $\hat{F}_I(t)$ and $\tilde{F}_I(t)$ are asymptotically unbiased and consistent.

For studying variance estimation with resampling methods, we assume that n/N is negligible, where $n = \sum n_h$, $N = \sum N_h$ and N_h is the number of first-stage clusters in the population.

3. A REPEATED HALF-SAMPLE BOOTSTRAP

When there are imputed missing data, naive bootstrap variance estimators obtained by treating the imputed data set, Y_I , as $Y = \{y_{hik} : (hik) \in s\}$, the data set of no missing values, do not capture the inflation in variance due to imputation and/or missing data and lead to serious underestimation. As a result, they are inconsistent. This is so, because simply treating Y_I as Y ignores the imputation process. This was noted by Shao and Sitter (1996) and they proposed re-imputing the bootstrap data set in the same way as the original data set was imputed. The bootstrap procedure in Shao and Sitter (1996) can be described as follows.

1. Draw a simple random sample $\{y_{hi}^* : i = 1, \dots, n_h - 1\}$ with replacement from the sample $\{\tilde{y}_{hi} : i = 1, \dots, n_h\}$, $h = 1, \dots, H$, independently across the strata, where $\tilde{y}_{hi} = \{y_{hij} : (h, i, j) \in s_r\} \cup \{\tilde{y}_{hij} : (h, i, j) \in s_m\}$.
2. Let a_{hij}^* be the response indicator associated with y_{hij}^* , $s_m^* = \{(h, i, j) : a_{hij}^* = 0\}$ and $s_r^* = \{(h, i, j) : a_{hij}^* = 1\}$. Apply the same imputation procedure used in constructing the imputed data set Y_I to the "nonrespondents" in s_m^* , using the "respondents" in s_r^* . Denote the bootstrap analogue of Y_I by Y_I^* .
3. Obtain the bootstrap analogue $\hat{\theta}_I^*$ of $\hat{\theta}$, based on the imputed bootstrap data set Y_I^* . For example, if $\hat{\theta} = \hat{Y}$ in (1) and $\hat{\theta}_I = \hat{Y}_I$ in (2), then

$$\hat{\theta}_I^* = \hat{Y}_I^* = \sum_{s_r^*} w_{hik}^* y_{hik}^* + \sum_{s_m^*} w_{hik}^* \tilde{y}_{hik}^*, \quad (4)$$

where \tilde{y}_{hik}^* is the imputed value using the bootstrap data and w_{hik}^* is $n_h/(n_h - 1)$ times the survey weight associated with y_{hik}^* (to reflect the fact that the bootstrap sample size is $n_h - 1$, not n_h). The bootstrap estimator of $\text{Var}(\hat{\theta}_I)$ is

$$v_B(\hat{\theta}_I) = \text{Var}^*(\hat{\theta}_I^*), \quad (5)$$

where Var^* is the conditional variance with respect to Y_I^* , given Y_I .

Shao and Sitter (1996) show that the bootstrap estimator defined in (5) is consistent for both smooth and nonsmooth estimators $\hat{\theta}$. When a random imputation method is considered, an implicit condition in their development is that $n_h/(n_h - 1)$ goes to 1. This can be seen from the special case of $\hat{\theta} = \hat{Y}$. From (2),

$$\begin{aligned} \text{Var}(\hat{Y}_I) &= \text{Var}\left[E_I(\hat{Y}_I)\right] + E\left[\text{Var}_I(\hat{Y}_I)\right] \\ &= \text{Var}\left(\frac{\sum_{s_r} w_{hik} y_{hik} \sum_s w_{hik}}{\sum_{s_r} w_{hik}}\right) + E\left(\hat{\sigma}^2 \sum_{s_m} w_{hik}^2\right), \end{aligned} \quad (6)$$

where

$$\hat{\sigma}^2 = \sum_{s_r} w_{hik} (y_{hik} - \bar{y}_r)^2 / \sum_{s_r} w_{hik},$$

$$\bar{y}_r = \sum_{s_r} w_{hik} y_{hik} / \sum_{s_r} w_{hik}.$$

Similarly, by (4),

$$\begin{aligned} \text{Var}^*(\hat{Y}_I^*) &= \text{Var}^*\left(\frac{\sum_{s_r^*} w_{hik}^* y_{hik}^* \sum_{s^*} w_{hik}^*}{\sum_{s_r^*} w_{hik}^*}\right) \\ &\quad + E^*\left(\hat{\sigma}^{*2} \sum_{s_m^*} w_{hik}^{*2}\right), \end{aligned} \quad (7)$$

where

$$\hat{\sigma}^{*2} = \sum_{s_r^*} w_{hik}^* (y_{hik}^* - \bar{y}_r^*)^2 / \sum_{s_r^*} w_{hik}^*,$$

$$\bar{y}_r^* = \sum_{s_r^*} w_{hik}^* y_{hik}^* / \sum_{s_r^*} w_{hik}^*.$$

From the theory of the bootstrap, the first terms on the right hand side of (6) and (7) converge to the same quantity, as do $\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$. Thus, Shao and Sitter's bootstrap is consistent if $\sum_{s_m} w_{hik}^2$ and $\sum_{s_m^*} w_{hik}^{*2}$ converge to the same quantity, which is true if $n_h/(n_h - 1)$ converges to 1 for all h , because

$$\begin{aligned} E^*\left(\sum_{s_m^*} w_{hik}^{*2}\right) &= E^*\left[\sum_{s^*} (1 - a_{hik}^*) w_{hik}^{*2}\right] \\ &= \sum_s (1 - a_{hik}) w_{hik}^2 n_h / (n_h - 1). \end{aligned}$$

The second term on the right hand side of (6) is the variance component corresponding to random imputation, which is typically a small portion of the overall variance. Thus, the overestimation due to $n_h/(n_h - 1)$ is serious only when the n_h 's are very small. The case $n_h = 2$ is, however, an important special case.

We now propose a bootstrap method which has no difficulty in the case of very small n_h 's while remaining valid more generally. Note that the use of bootstrap sample size $n_h - 1$ is to ensure that the first term on the right hand side of (7) has the same limit as the first term on the right

hand side of (6) (Rao and Wu 1988). When n_h is used as the bootstrap sample size in stratum h , Rao and Wu (1988) showed that in the case of no missing data, the bootstrap variance estimator underestimates. They proposed a rescaling to circumvent the problem, but rescaling does not produce correct bootstrap estimators in the presence of imputed data.

What is ideally required for our problem is a bootstrap method with the bootstrap sample size equal to the original sample size n_h which produces an asymptotically unbiased variance estimator (in the case of no missing data) without rescaling. We now show that this can be accomplished as follows. Suppose that there is no missing data and that all of the $n_h = 2m_h$'s are even. Take a simple random sample of size m_h without replacement independently from $\{y_{hi}: i = 1, \dots, n_h\}$ and repeat each obtained unit twice to get $\{y_{hi}^*: i = 1, \dots, n_h\}$. We call this method the repeated half-sample bootstrap. The resulting v_B will then be approximately unbiased and consistent. In the linear case where $\hat{Y} = \sum_{(hik)} w_{hik} y_{hik} = \sum_h \sum_{i=1}^{n_h} y_{hi} / n_h = \sum_h \bar{y}_h$ and $y_{hi} = \sum_{k=1}^{n_{hi}} n_h w_{hik} y_{hik}$, the consistency of v_B follows from

$$\begin{aligned} \text{Var}^*(\hat{Y}^*) &= \sum_h \text{Var}^*(\bar{y}_h^*) = \sum_h \text{Var}^*\left(\frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}^*\right) \\ &= \sum_h \text{Var}^*\left(\frac{2m_h}{n_h} \frac{1}{m_h} \sum_{i=1}^{m_h} y_{hi}^*\right) \\ &= \sum_h \text{Var}^*\left(\frac{1}{m_h} \sum_{i=1}^{m_h} y_{hi}^*\right) \\ &= \sum_h \frac{(1-1/2)}{m_h} \frac{1}{n_h-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \\ &= \sum_h s_h^2 / n_h, \end{aligned}$$

the usual approximately unbiased and consistent estimator of variance, where $s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$. The consistency of v_B for a nonlinear $\hat{\theta}_I$ follows from the linear case and Taylor's expansion, when $\hat{\theta}_I$ is a function of weighted averages, or the arguments used in Shao and Rao (1994), Shao and Sitter (1996), and Shao *et al.* (1998) when $\hat{\theta}_I$ is non-smooth such as a median.

If $n_h = 2m_h + 1$ is odd, it is not possible to take an exact half-sample. In this case, the following two results lead us to an adaptation of the above idea:

- i) If we choose a simple random resample of size $m_h = (n_h - 1)/2$ without replacement and repeat each unit twice, we end up with $n_h - 1$ units. If we obtain an additional unit by selecting one at random from the $n_h - 1$ units already resampled, $\text{Var}^*(\hat{Y}^*) = \sum_h (n_h + 3) s_h^2 / n_h^2$;

- ii) If we choose a simple random resample of size $m_h + 1$ without replacement and repeat each unit twice, we end up with $n_h + 1$ units. If we discard one of these at random, $\text{Var}^*(\hat{Y}^*) = \sum_h (n_h - 1) s_h^2 / n_h^2$.

Thus, if we used method (i) with probability 1/4 and method (ii) with probability 3/4 at each bootstrap replication, we obtain the desired result. This repeated half-sample bootstrap method yields approximately unbiased variance estimates without rescaling and has a bootstrap sample size equal to the original sample size. Thus, if we use this bootstrap for Step 1 of the method of Shao and Sitter (1996) as described above, the resulting bootstrap estimators are asymptotically unbiased and consistent for any n_h , under the regularity conditions stated in Shao and Sitter (1996) and Shao *et al.* (1998).

4. THE PROPER MONTE CARLO FOR THE BOOTSTRAP

If v_B in (5) has no explicit form, one may use the Monte Carlo approximation

$$v_B(\hat{\theta}_I) \approx \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_{I(b)}^* - \bar{\theta}_I^*)^2, \quad (8)$$

where $\bar{\theta}_I^* = B^{-1} \sum_{b=1}^B \hat{\theta}_{I(b)}^*$, $\hat{\theta}_{I(b)}^* = \hat{\theta}(Y_{I(b)}^*)$, and $Y_{I(b)}^*$, $b = 1, \dots, B$, are independent re-imputed bootstrap data sets. It is common practice in many applications of the bootstrap to replace the average of the bootstrap estimators $\bar{\theta}_I^*$ in (8) by the original estimator $\hat{\theta}_I$ (see Rao and Wu 1985, page 232). The latter is simpler to use and is thus the most common. With no imputed data, this is usually correct. However, using the analogue with the re-imputed bootstrap is not correct. The reason is that $\hat{\theta}_I$ is the result of a single realization of the random imputation, while $\bar{\theta}_I^* \approx E^*(\hat{\theta}_I^*) \approx E_I(\hat{\theta}_I)$ since we are averaging over repeated re-imputations, and $\hat{\theta}_I$ and $E_I(\hat{\theta}_I)$ are not close for random imputation. When $\hat{\theta}_I = \hat{Y}_I$, for example, $E_I(\hat{Y}_I) = \hat{Y}_I$ given in section 2 and the difference $\hat{Y}_I - \hat{Y}_I$ is not a relatively negligible term when random imputation is used. Thus,

$$v_{B2} = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_{I(b)}^* - \hat{\theta}_I)^2 = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_{I(b)}^* - \bar{\theta}_I^*)^2 + (\bar{\theta}_I^* - \hat{\theta}_I)^2$$

and the first term goes to $\text{Var}^*(\hat{\theta}_I^*)$ as $B \rightarrow \infty$ but the second term does not go to zero which implies that v_{B2} badly overestimates the variance. This is not only true for the proposed repeated half-sample bootstrap but also for those considered in Shao and Sitter (1996).

One should also note that using the $\hat{\theta}_{I(b)}^*$, $b = 1, \dots, B$ to obtain bootstrap confidence intervals via the percentile method avoids this concern since the histogram of these values will be correctly centered about $E^*(\hat{\theta}_I^*)$. However, one must take more care with bootstrap- t confidence

intervals. It is important that one define $t_b^* = (\hat{\theta}_{I(b)}^* - \bar{\theta}_{I(\cdot)}^*)/\sigma_b^*$ (not $t_b^* = (\hat{\theta}_I^* - \hat{\theta}_I)/\sigma_b^*$) and use $\{\hat{\theta}_I - t_U \sigma_b^*, \hat{\theta}_I - t_L \sigma_b^*\}$, where $\sigma_b^{*2} = v_B(\mathbf{Y}_I^*)$, $t_L^* = \text{CDF}_I^{-1}(\alpha)$, $t_U^* \text{CDF}_I^{-1}(1 - \alpha)$ and $\text{CDF}_I(x) = \#\{t_b^* \leq x; b = 1, \dots, B\}/B$.

5. A REPEATED BRR

We first describe the most common application of the BRR, $n_h = 2$ clusters per stratum (McCarthy 1969) in the setting of no missing data. A set of B balanced half-samples or replicates is formed by deleting one first-stage cluster from the sample in each stratum, where this set is defined by a $B \times H$ matrix $(\delta_{bh})_{B \times H}$ with $\delta_{bh} = +1$ or -1 according to whether the first or the second first-stage cluster of stratum h is in the b -th half-sample and $\sum_{b=1}^B \delta_{bh} \delta_{bh'} = 0$ for all $h \neq h'$; that is, the columns of the matrix are orthogonal. A minimal set of B balanced half-samples can be constructed from a $B \times B$ Hadamard matrix by choosing any H columns excluding the column of all $+1$'s, where $H + 1 \leq B \leq H + 4$. Let $\hat{\theta}_{(b)}$ be the survey estimator computed from the b -th half-sample. The estimator $\hat{\theta}_{(b)}$ can be obtained using the same formula as for $\hat{\theta}$ with w_{hik} changed to $w_{hik(b)}$, which equals $2w_{hik}$ or 0 according to whether or not the (hi) -th cluster is selected in the b -th half-sample or not. The BRR variance estimator for $\hat{\theta}$ is then given by

$$v_{\text{BRR}} = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_{(b)} - \bar{\theta}_{(\cdot)})^2, \quad (9)$$

where $\bar{\theta}_{(\cdot)} = \sum_b \hat{\theta}_{(b)}/B$, and is often replaced by $\hat{\theta}$. The variance estimator v_{BRR} has been shown to be consistent for smooth functions of estimated totals by Krewski and Rao (1981) and for nonsmooth estimators by Shao, and Wu (1992) and Shao and Rao (1994).

A naive BRR for problems with randomly imputed data would be obtained as in (9) with $\hat{\theta}_{(b)}$ and $\bar{\theta}_{(\cdot)}$ replaced by $\hat{\theta}_{I(b)}$ and $\bar{\theta}_{(\cdot)} = B^{-1} \sum_b \hat{\theta}_{I(b)}$, where $\hat{\theta}_{I(b)}$ is the estimator calculated from \mathbf{Y}_I using the BRR weights. But this produces inconsistent variance estimators because it fails to take into account the effect of missing data and the random imputation.

To correctly apply the BRR in the presence of random imputation by using re-imputation, we must deal with the issue of n_h being small. Recall that for the bootstrap such small n_h 's caused difficulty because the stratum resample size, $n_h - 1$, was smaller than the original stratum sample size, n_h . This is true for the BRR, as well. We propose an easy way to circumvent this difficulty. Rather than obtaining the b -th BRR replicate of the estimator, $\hat{\theta}_{(b)}$, from the same formula as for $\hat{\theta}$ but with weights $w_{hik(b)}$ equal $2w_{hik}$ or 0 according as to whether the (hi) -th cluster is selected in the b -th half-sample or not, instead use the original weights but include the (hi) -th cluster twice or not at all according as to whether the (hi) -th cluster is selected

in the b -th half-sample or not. If we view the BRR in this way: i) the resulting v_{BRR} in (9) remains the same; and ii) the resample size is the same as the original sample size. This repeated BRR can be viewed as a type of balanced bootstrap, however one should note that the balanced bootstrap described in Nigam and Rao (1996) for the case of no missing data does not work in this case because, though it uses a resample size $n_h = 2$ in each stratum, it does so in such a way as to still require rescaling and thus will not work in the presence of random imputation.

The proposed repeated BRR has no difficulty in the presence of random imputation. The procedure becomes

1. Form the set of half-samples, 1 unit per stratum, using a Hadamard matrix as described above.
2. Obtain the b -th BRR replicate by repeating each unit in the obtained half-sample twice. Denote this $\{y_{hi}^*; i = 1, \dots, n_h = 2\}$.
3. Let a_{hij}^* be the response indicator associated with $y_{hij}^*, s_m^* = \{(h, i, j): a_{hij}^* = 0\}$, and $s_r^* = \{(h, i, j): a_{hij}^* = 1\}$. Apply the same imputation procedure used in constructing \mathbf{Y}_I to the units in s_m^* , using the "respondents" in s_r^* . Denote the b -th BRR replicate of \mathbf{Y}_I by $\mathbf{Y}_{I(b)}^*$.
4. Obtain the BRR analogue $\hat{\theta}_{I(b)}^*$ of $\hat{\theta}$, based on the imputed BRR data set $\mathbf{Y}_{I(b)}^*$.
4. Repeat 1-4 for each row of the $B \times H$ matrix to get $\hat{\theta}_{I(b)}^*$ for $b = 1, \dots, B$ and apply the standard BRR formula (9) to obtain BRR variance estimators for $\hat{\theta}_I$, with $\bar{\theta}_{(\cdot)} = B^{-1} \sum_b \hat{\theta}_{I(b)}^*$ (For the same reason that is discussed in section 4, we should not replace $\bar{\theta}_{I(\cdot)}$ by $\hat{\theta}_I$).

We can extend this idea to cases with $n_h > 2$ by using the same strategy with half-samples obtained from balanced orthogonal multi-arrays (BOMA's) (Sitter 1993). For example, Table 1 gives a set of $B = 24$ balanced resamples for $H = 7$ strata with $n_h = 4$ psu's in each stratum. It is derived using the BOMA given in Table 1 of Sitter (1993) and repeating each resampled unit twice as in Step 2 above. Using a BOMA in Steps 1 and 2 of the procedure above also results in an approximately unbiased variance estimator. BOMA's are fairly easily constructed for even n_h using balanced incomplete block designs and Hadamard matrices, but are difficult to construct for odd n_h . They can also handle unequal n_h 's for different strata, though construction becomes a more serious problem (see Sitter 1993).

6. A SIMULATION

To study the properties of the proposed resampling variance estimators, we consider a finite population of $H = 32$ strata with N_h clusters in stratum h and ten ultimate units in each cluster. The characteristic of interest y_{hik} are generated as follows:

Table 1
A Set of Balanced Resamples Constructed from a BOMA

b	h						
	1	2	3	4	5	6	7
1	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)
2	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)
3	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
4	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
5	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
6	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
7	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)
8	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)
9	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)
10	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)
11	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)
12	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)
13	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)
14	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)
15	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)
16	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)
17	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)
18	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)
19	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)
20	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)
21	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)
22	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)
23	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)
24	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)

$$y_{hik} = y_{hi} + \varepsilon_{hik},$$

where $y_{hi} \sim N(\mu_h, \sigma_h^2)$ independent of $\varepsilon_{hik} \sim N(0, [1 - \rho]\sigma_h^2/\rho)$ and the parameter values are those given in Table 2. For a particular value of the intracluster correlation, ρ , a single finite population was thus generated and then fixed and repeatedly sampled from. Each simulation consisted of selecting $n_h = 2$ clusters with replacement from stratum h for $h = 1, \dots, H$ and enumerating the entire cluster. Each ultimate unit in the obtained cluster was independently declared a respondent or nonrespondent with probability p and $(1 - p)$ respectively, *i.e.*, uniform response. The nonrespondents were then imputed both using random imputation and adjusted random imputation and the population total and distribution function, for various values of $F(t)$, were estimated. Two values of ρ , 0.1 and 0.3, and two values of p , 0.6 and 0.8, were considered. Note that the first-stage sampling fraction is quite small (0.064), so that with-replacement and without-replacement sampling are essentially equivalent.

To compare the performance of the different variance estimators we calculated the percent relative bias and relative instability for each, defined as

$$\% \text{RB} = \frac{100}{S} \sum_{s=1}^S y_s(\hat{\theta}_I) / \text{MSE}(\hat{\theta}_I)$$

and

$$\text{RI} = \left\{ \frac{1}{S} \sum_{s=1}^S [v_s(\hat{\theta}_I) - \text{MSE}(\hat{\theta}_I)]^2 \right\}^{1/2} / \text{MSE}(\hat{\theta}_I),$$

respectively, where the number of simulation runs was $S = 5,000$ and the true $\text{MSE}(\hat{\theta}_I)$ was obtained through an independent set of 50,000 simulation runs. The bootstrap variance estimators were each based on $B = 2,000$ bootstrap resamples. We obtain results for estimating the variance of $\hat{\theta}_I$ equal to the imputed total and the imputed distribution function using: (i) the repeated half-sample bootstrap with proper Monte Carlo approximation, v_B , as in equation (8) and with improper Monte Carlo approximation replacing $\bar{\theta}_{I(\cdot)}^*$ with $\hat{\theta}_I$ denoted v_{B2} ; and (ii) the proper repeated BRR, v_{BRR} , as in equation (9) and the improper repeated BRR replacing $\bar{\theta}_{I(\cdot)}$ with $\hat{\theta}_I$ denoted $v_{\text{BRR}2}$.

Table 3 summarizes the results for percent relative bias using random imputation and adjusted random imputation. Note that adjusted random imputation is not presented for estimating the population total, Y , as adjusted random imputation removes the imputation variance from the estimator and thus simpler methods of variance estimation are available (Chen *et al.* 2000). It is clear from the high %RB for v_{B2} and $v_{\text{BRR}2}$ that one must not replace $\bar{\theta}_{I(\cdot)}$ and $\bar{\theta}_{I(\cdot)}^*$ by $\hat{\theta}_I$ in the bootstrap or the BRR, respectively. It is also clear that both the repeated half-sample bootstrap and the repeated BRR variance estimators, v_B and v_{BRR} have negligible bias when properly applied.

Table 2
Parameters of the Finite Population

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	200	20.0	17	31	150	15.0
2	16	175	17.5	18	31	140	14.0
3	20	150	15.0	19	31	130	13.0
4	25	190	19.0	20	34	120	12.0
5	25	165	16.5	21	34	110	11.0
6	25	190	19.0	22	34	100	10.0
7	25	180	18.0	23	34	150	15.0
8	28	170	17.0	24	37	125	12.5
9	28	160	16.0	25	37	100	10.0
10	28	180	18.0	26	37	150	15.0
11	31	170	17.0	27	37	125	12.5
12	31	160	16.0	28	39	100	10.0
13	31	150	15.0	29	39	75	7.5
14	31	180	18.0	30	42	75	7.5
15	31	170	17.0	31	42	75	7.5
16	31	160	16.0	32	42	75	7.5

Given the results of Table 3, we consider relative instability, RI, only for v_B and v_{BRR} . We also restrict our presentation to $\rho = 0.3$ and $p = 0.6$ as the RI results were qualitatively the same in the other three cases. These results are given in Table 4. As one can see, though the differences are small, v_B is slightly more stable than v_{BRR} . This was generally the case for all values of ρ and p . We also included the adjusted jackknife of Rao and Shao (1992) and the adjusted BRR of Shao *et al.* (1998) in simulations for $\theta = Y$ and v_B again was uniformly more stable. For example, with $\rho = 0.3$ and $p = 0.6$ as in Table 4, RI for the adjusted jackknife and the adjusted BRR were both 0.27. This may be because the reimputation approach has an advantage in estimating the component of the variance due to the imputation against the adjustment approach, provided the resample size is large enough to eliminate Monte Carlo error as is the case in our simulations. But, when the number of reimputations is moderate (like in the BRR with reimputation or the bootstrap with $B = 1,000$), this advantage is not entirely realized.

Table 3
% RB for v_B, v_{B2}, v_{BRR} and v_{BRR2}

Estimand	Random imputation				Adjusted random imputation			
	v_{BRR}	v_{BRR2}	v_B	v_{B2}	v_{BRR}	v_{BRR2}	v_B	v_{B2}
$\rho = 0.1$ and $p = 0.6$								
Y	0.00	21.54	0.79	21.60				
F(t) = 0.0625	-1.09	15.92	-0.52	15.88	0.46	19.64	1.24	19.51
F(t) = 0.2500	-0.13	19.44	0.62	19.55	0.85	14.86	1.80	15.08
F(t) = 0.5000	-0.36	21.68	0.52	21.55	0.55	10.73	1.24	10.76
F(t) = 0.7500	-0.84	19.89	0.13	20.09	-0.36	10.98	0.54	11.31
F(t) = 0.9375	0.05	21.92	0.57	21.66	0.81	19.12	1.39	18.91
$\rho = 0.1$ and $p = 0.8$								
Y	-0.63	15.06	0.36	15.37				
F(t) = 0.0625	-1.99	10.30	-1.72	10.16	-1.65	10.97	-1.08	11.13
F(t) = 0.2500	-1.27	13.65	-0.88	13.30	-0.95	8.89	-0.52	8.81
F(t) = 0.5000	-0.72	15.26	0.02	15.26	-0.12	6.58	0.25	6.53
F(t) = 0.7500	-0.37	14.50	0.57	14.76	0.36	7.56	1.05	7.81
F(t) = 0.9375	-0.14	16.16	0.75	16.36	0.56	13.04	1.22	13.08
$\rho = 0.3$ and $p = 0.6$								
Y	0.25	21.34	0.78	21.09				
F(t) = 0.0625	-1.39	11.45	-0.86	11.37	-0.35	15.38	0.64	15.64
F(t) = 0.2500	-0.41	19.89	0.14	19.73	1.23	13.79	1.71	13.62
F(t) = 0.5000	-0.10	20.25	0.37	19.89	0.29	8.97	0.78	8.88
F(t) = 0.7500	-1.40	16.70	-0.49	16.89	-0.75	9.24	0.07	9.49
F(t) = 0.9375	0.71	17.78	1.03	17.57	0.91	15.07	1.34	15.04
$\rho = 0.3$ and $p = 0.8$								
Y	0.01	15.22	0.93	15.51				
F(t) = 0.0625	-1.09	7.54	-0.56	7.69	-1.24	8.64	-0.35	9.07
F(t) = 0.2500	-0.44	15.22	-0.08	14.99	-0.23	8.18	0.29	8.23
F(t) = 0.5000	0.05	14.92	0.71	14.84	0.43	6.21	0.86	6.20
F(t) = 0.7500	0.13	12.54	0.86	12.70	0.81	6.85	1.26	6.99
F(t) = 0.9375	1.62	13.13	2.06	13.01	1.86	11.04	2.34	11.02

Table 4
RI for v_B and v_{BRR} with $\rho = 0.3$ and $p = 0.6$

Estimand	Random imputation		Adjusted random imputation	
	v_{BRR}	v_B	v_{BRR}	v_B
Y	0.27	0.23		
$F(t) = 0.0625$	0.60	0.59	0.57	0.56
$F(t) = 0.2500$	0.35	0.32	0.37	0.35
$F(t) = 0.5000$	0.27	0.23	0.28	0.26
$F(t) = 0.7500$	0.29	0.26	0.30	0.28
$F(t) = 0.9375$	0.48	0.46	0.48	0.46

7. CONCLUSION

We proposed repeated half-sample bootstrap and balanced repeated replication methods for variance estimation in the presense of random imputation that capture the imputation variance by reimputing for each replication using the same random imputation method as in the original sample. These repeated half-sample methods are valid in stratified multi-stage sampling, even when the number of psu's sampled in each stratum is very small, *e.g.*, 2. The key is that these methods use a stratum resample size that is equal to the original sample size without resorting to rescaling. These provide a unified method that works irrespective of the imputation method (random or non-random), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). It is important to note that using reimputation to capture the imputation variance requires that one take greater care in the definition of the BRR and the Monte Carlo approximation to the bootstrap variance. In both cases it is important to use the mean of the replicates in the definition as opposed to replacing it with the estimator applied to the original sample.

ACKNOWLEDGEMENTS

Hiroshi Saigo was supported by grants from the Promotion and Mutual Aid Corporation for Private Universities of Japan and the Japan Economic Research Foundation. Jun Shao was supported by National Science Foundation Grant DMS-0102223, and National Security Agency Grant MDA904-99-1-0032. Randy R. Sitter was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank all referees for their helpful comments and suggestions.

REFERENCES

CHEN, J., RAO, J.N.K. and SITTER, R.R. (2000). Adjusted imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

EFRON, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89, 463-479.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

MANTEL, H.J., and SINGH, A.C. (1991). Standard errors of estimates of low proportions: A proposed methodology. Technical Report, Statistics Canada.

MCCARTHY, P.J. (1969). Pseudoreplication half samples. *Review of the International Statistical Institute*, 37, 239-264.

NIGAM, A.K., and RAO, J.N.K. (1996). On balanced bootstrap, for stratified multistage samples. *Statistica Sinica*, 6, 199-214.

RAO, J.N.K. (1993). Linearization variance estimators under imputation for missing data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for non-linear statistics. *Journal of the American Statistical Association*, 80, 620-630.

RAO, J.N.K., and WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RUBIN, D.B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

SHAO, J., and RAO, J.N.K. (1994). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā, B*, Special Volume 55, 393-414.

SHAO, J., and SITTER, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

SHAO, J., and WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics*, 20, 1571-1593.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

SITTER, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Local Polynomial Regression in Complex Surveys

D.R. BELLHOUSE and J.E. STAFFORD¹

ABSTRACT

Local polynomial regression methods are put forward to aid in exploratory data analysis for large-scale surveys. The proposed method relies on binning the data on the x -variable and calculating the appropriate survey estimates for the mean of the y -values at each bin. When binning on x has been carried out to the precision of the recorded data, the method is the same as applying the survey weights to the standard criterion for obtaining local polynomial regression estimates. The alternative of using classical polynomial regression is also considered and a criterion is proposed to decide whether the nonparametric approach to modeling should be preferred over the classical approach. Illustrative examples are given from the 1990 Ontario Health Survey.

KEY WORDS: Covariates; Exploratory data analysis; Kernel smoothing; Regression.

1. INTRODUCTION

Following Fuller (1975), multiple linear regression techniques have been studied and used extensively in sample surveys. At least three chapters of Skinner, Holt and Smith (1989) are devoted to this subject. Here we restrict attention to the case in which there is one covariate x for the variate of interest y so that we could consider polynomial regression as well as simple linear regression. In this context we could also consider the nonparametric approach of local polynomial regression, which, for the case of independent and identically distributed random variables, is described in Hardle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996) and Eubank (1999). Using the survey weights, Korn and Graubard (1998) introduced the use of local polynomial regression for graphical display of complex survey data. However, they did not provide any statistical properties for their procedures. Smith and Njenga (1992) used regression kernel smoothing techniques to obtain robust estimates of the mean and regression parameters for an assumed superpopulation model. Here we use local polynomial regression as an exploratory tool to discover relationships between y and its covariate x .

We assume that the covariate x is measured on a continuous scale. Due to the precision at which the data are recorded for the survey file and the size of the sample, there will be multiple observations at many of the distinct values. This feature of large-scale survey data has been exploited by Hartley and Rao (1968, 1969) in their scale-load approach to the estimation of finite population parameters. Here we exploit this same feature of the data to examine the relationship between y and its covariate x . In recognizing that the data may be naturally binned to the precision of the data, we can consider taking a further step by constructing larger bin sizes. Under this approach we examine the effect

of the sampling design on estimates and second order moments.

Suppose that in the finite population of size N , x has k distinct values so that natural binning has taken place, or that x has been categorized into k bins that are wider than the precision of the data. Let x_i be the value of x representing the i^{th} bin, and assume that the values of x_i are equally spaced. The spacing or bin size $b = x_i - x_{i-1}$. The finite population mean for the y -values at x_i is \bar{y}_i . We assume that a sample of size n taken from this population has the same structure as the population in that there are k bins. From the sample data we calculate the survey estimate of \bar{y}_i of \bar{y}_i . The finite population proportion of the observations with value x_i is denoted by p_i . This proportion is estimated by the survey estimate \hat{p}_i . We assume that \hat{y}_i and \hat{p}_i are asymptotically unbiased, in the sense of Särndal, Swensson and Wretman (1992, pages 166-167), for \bar{y}_i and p_i respectively. The survey estimates \hat{y}_i for $i = 1, \dots, k$ have variance-covariance matrix \mathbf{V} . On considering the distinct values x_i as domains, the estimated variance-covariance matrix $\hat{\mathbf{V}}$ may be obtained easily through survey packages such as SUDAAN and STATA.

There are several advantages to binning the data on the covariate x for exploratory data analysis:

- For large surveys, a plot of \hat{y}_i against x_i may be more informative and less cluttered than a plot of the raw data.
- By appealing to a finite population central limit theorem on \hat{y}_i and imposing a superpopulation assumption on \bar{y}_i , a relatively simple model for \hat{y}_i may be assumed so that the analyst may easily focus on the central issue considered here, determination of the trend function in x .

¹ D.R. Bellhouse Department of Statistical and Actuarial Sciences, Western Science Centre, University of Western Ontario, London, Ontario N6A 5B7, e-mail: bellhouse@stats.uwo.ca; J.E. Stafford, Department of Public Health Sciences, Faculty of Medicine, McMurich Building, University of Toronto, Toronto, Ontario, M5S 1A8, e-mail: stafford@utstat.toronto.edu.

- Once \hat{V} has been obtained, then a wide variety of powerful exploratory data analyses can be easily carried out in languages such as S-Plus. Working with the raw data requires continued appeals to SUDAAN or STATA for the appropriate variance estimates.
- By binning the data, an approach to regression analysis is obtained that provides a parallel to other nonparametric approaches to survey data analysis. For example, in categorical data analysis obtained initially by Rao and Scott (1981), in the logistic regression approach of Roberts, Rao and Kumar (1987) or in the generalized linear model approach of Bellhouse and Rao (2000), the tests and associated distributions are obtained through survey estimates of domain means or proportions.

For the superpopulation, we assume that we have a model such that $E_m(\bar{y}_i) = m(x_i)$, where E_m is the superpopulation expectation. We assume further that as we move to a continuum of values on x , then $m(x)$ is a smooth function. The function $m(x)$ is the ultimate function of interest for estimation. In section 2 we provide local polynomial regression methods to estimate $m(x)$. These methods are applied to data from the 1990 Ontario Health Survey in section 3. In section 4, the question is asked: would the classical polynomial regression techniques have served equally as well in modeling $m(x)$? Some future directions for this work are given in section 5. Generally, we adopt the notation of Wand and Jones (1995) in discussing local polynomial regression here.

2. BASIC METHODOLOGY

For local polynomial regression, the nestimate of $m(x)$ at any value of x is obtained upon minimizing

$$\sum_{i=1}^k \hat{p}_i \left\{ \hat{\bar{y}}_i - \beta_0 - \beta_1(x_i - x) - \dots - \beta_q(x_i - x)^q \right\}^2 K((x_i - x)/h) \quad (1)$$

with respect to $\beta_0, \beta_1, \dots, \beta_q$. The values that minimize (1) are denoted by $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_q$. Further, for the given value of x , $\hat{m}(x) = \hat{\beta}_0$. In (1), the kernel $K(t)$ is a symmetric function with $\int K(t) dt = 1$, $\int t K(t) dt = 0$, $0 < \int t^2 K(t) dt < \infty$ and

$$R(K) = \int [K(t)]^2 dt < \infty. \quad (2)$$

Also in (1), h is the window width of the kernel. In minimizing (1) to obtain local polynomial regression estimates, there are two possibilities for binning on x . The first is to bin to the precision of the recorded data so that $\hat{\bar{y}}_i$ is calculated at each distinct outcome of x . In other situations it may be practical to pursue a binning on x that is rougher than the accuracy of the data.

In moving from the sample to the population we maintain the same window width h . This is in contrast to Breidt and Opsomer (2000) and Buskirk (1999) who assume a smoothing parameter h_N for smoothing in the full finite population. In the context here, this would yield a function $m_N(x)$, the finite population smoothed version of the \bar{y}_i with smoothing parameter h_N , as a finite population parameter of interest followed by $m(x)$ the hypothetical smooth function under the asymptotic assumptions. We have kept h constant in view of the way in which binning that has been done; the bin structure is the same in the sample as in the population. The choice of the smoothing parameter h depends on the spacing of the x 's and the variation in the data (Green and Silverman 1994, pages 43–44). The spacing of the covariate is usually dominant in the determination of h . Since the spacing has been kept constant from sample to finite population with the spacing changing only when the asymptotic assumptions are applied, we keep $h_N = h$.

Korn and Graubard (1998) provide a slightly different objective function to (1). They replace the sum over the bins in (1) by the sum over all sampled units and \hat{p}_i in (1) by the sample weights. Korn and Graubard's objective function reduces to (1) plus a term that involves the weighted sum of squares of deviations of sample observations from the binned means where the weights are the sample weights scaled to sum to one. Consequently, the estimate of $m(x)$ is the same in both cases.

The estimate $\hat{m}(x)$ and its first two moments can be expressed in matrix notation. The forms are exactly the same as those that appear, for example, in Wand and Jones (1995, chapter 5.3) whose notation we have adopted. Let the vector of finite population means at the distinct values of x be $\bar{\mathbf{y}} = (\bar{y}_1, \dots, \bar{y}_k)^T$ and let $\hat{\mathbf{y}}$ be its vector of survey estimates. Further, let

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \dots & (x_1 - x)^q \\ 1 & x_2 - x & \dots & (x_2 - x)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_k - x & \dots & (x_k - x)^q \end{bmatrix}$$

and

$$\mathbf{W}_x = \frac{1}{h} \text{diag} \left(p_1 K((x_1 - x)/h), \dots, p_k K((x_k - x)/h) \right).$$

The matrix $\hat{\mathbf{W}}_x$ is \mathbf{W}_x with p replaced by \hat{p} . Then

$$\hat{m}(x) = \mathbf{e}^T (\mathbf{X}_x^T \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \hat{\mathbf{W}}_x \hat{\mathbf{y}}, \quad (3)$$

where \mathbf{e} is the $k \times 1$ vector $(1, 0, 0, \dots, 0)^T$. The approximate design-based expectation of $\hat{m}(x)$ is

$$E_p(\hat{m}(x)) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \bar{\mathbf{y}}, \quad (4)$$

where E_p denotes expectation with respect to the sampling design. We can also consider (4) as a smoothed estimate of $m(x)$ so that $\hat{m}(x)$ is also an estimate of $m(x)$. In the derivation of (4) we note that $E_p(\hat{\mathbf{y}}) = \bar{\mathbf{y}}$ and $E_p(\hat{\mathbf{W}}_x) = \mathbf{W}_x$ for large sample size n . Further, in (3) we can write $\hat{\mathbf{W}}_x = \mathbf{W}_x + \hat{\mathbf{A}}$, where $\hat{\mathbf{A}} = \hat{\mathbf{W}}_x - \mathbf{W}_x$. We use the first two terms in the expansion $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \dots$ as an approximation to complete the derivation. Using the same techniques, the approximate design-based variance is given by

$$V_p(\hat{m}(x)) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{V} \mathbf{W}_x \mathbf{X}_x (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{e}. \tag{5}$$

The results in (4) and (5) were obtained ignoring higher order terms in $1/n$. An estimate of the variance $\hat{V}_p(\hat{m}(x))$ is obtained on substituting the survey estimate $\hat{\mathbf{V}}$ for \mathbf{V} and $\hat{\mathbf{W}}_x$ for \mathbf{W}_x in (5).

3. EXAMPLES FROM THE ONTARIO HEALTH SURVEY

We illustrate local polynomial regression techniques with data from the Ontario Health Survey (Ontario Ministry of Health 1992). This survey was carried out in 1990 using a stratified two-stage cluster sample. The purpose was to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of morbidity and mortality in Ontario. The survey was designed to be compatible with the Canada Health Survey carried out in 1978-79. A total sample size of 61,239 people was obtained from 43 public health units across Ontario. The public health unit was the basic stratum with an additional division of the health unit into rural and urban strata so that there were a total of 86 strata. The first stage units within a stratum were enumeration areas taken from the 1986 Census of Canada. An average of 46 enumeration areas was chosen within each stratum. Within an enumeration area, dwellings were selected, approximately 15 from an urban enumeration area and 20 from a rural enumeration area. Information was collected on members of the household within the dwelling.

Several health characteristics were measured. We focus on one continuous variable from the survey, Body Mass Index (BMI). The BMI is a measure of weight status and is calculated from the weight in kilograms divided by the square of the height in meters. The index is not applicable to adolescents, adults over 65 years of age and pregnant or breastfeeding women. The measure varies between 7.0 and 45.0. A value of the BMI less than 20.0 is often associated with health problems such as eating disorders. An index value above 27.0 is associated with health problems such as hypertension and coronary heart disease. Associated with

the BMI is another measure, the Desired Body Mass Index (DBMI). The DBMI is the same measure as BMI with actual weight replaced by desired weight. A total of 44,457 responses were obtained for the BMI and 41,939 for the DBMI.

When there are only a few distinct outcomes of x , binning on x is done in a natural way. For example, in investigating the relationship between the body mass index (BMI) and age, the age of the respondent was reported only at integral values. The solid dots in Figure 1 are the survey domain estimates of the average BMI (\hat{y}_i) for women at each of the ages 18 through 65 (x_i). The solid and dotted lines show the plot of $\hat{m}(x)$ against x using bandwidths $h = 7$ and $h = 14$ respectively. It may be seen from Figure 1 that BMI increases approximately linearly with age until around age 50. The increase slows in the early 50s, peaks at age 55 or so, and then begins to decrease. On plotting the trend lines only for BMI and the desired body mass index (DBMI) for females as shown in Figure 2, it may be seen that, on average, women desire to reduce their BMI at every age by approximately two units.

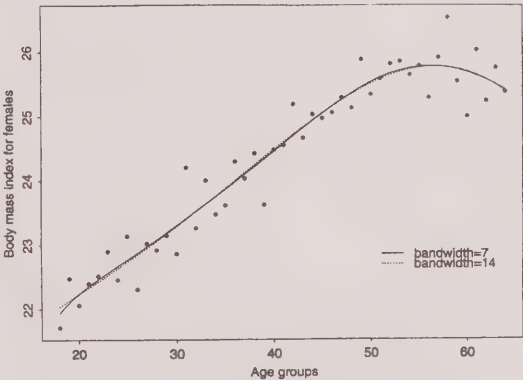


Figure 1. Age trend in BMI for females

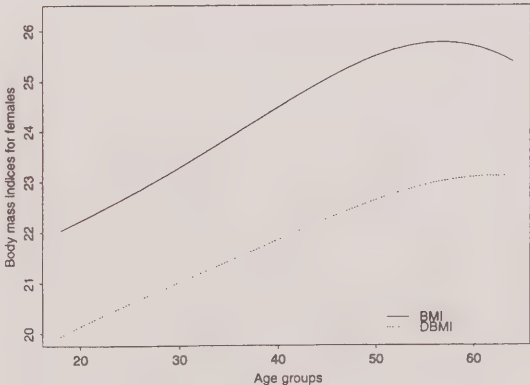


Figure 2. Age trends for females

In other situations it is practical to construct bins on x wider than the precision of the data. To investigate the relationship between what women desire for their weight ($\text{DBMI} = \hat{y}_i$) and what women actually weigh ($\text{BMI} = x_i$) the x -values were grouped. Since the data were very sparse for values of BMI below 15 and above 42, these data were removed from consideration. The remaining groups were 15.0 to 15.2, 15.3 to 15.4 and so on, with the value of x_i chosen as the middle value in each group. The binning was done in this way for the purposes of illustration to obtain a wide range of equally spaced nonempty bins. For each group the survey estimate \hat{y}_i was calculated. The solid dots in Figure 3 show the survey estimates of women's DBMI for each grouped value of their respective BMI. The scatter at either end of the line reflects the sampling variability due to low sample sizes. The plot shows a slight desire to gain weight when the BMI is at 15. This desire is reversed by the time the BMI reaches 20 and the gap between the desire (DBMI) and reality (BMI) widens as BMI increases.

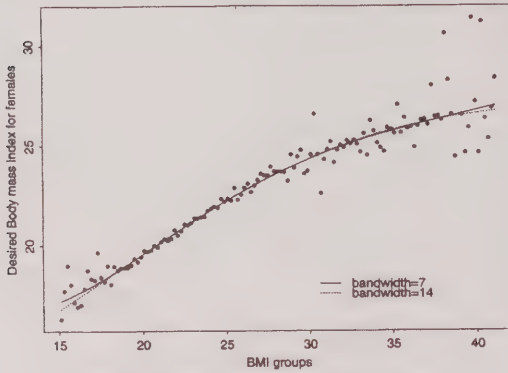


Figure 3. BMI trend in DBMI for females

4. PARAMETRIC VERSUS NONPARAMETRIC REGRESSION

Local polynomial regression allows us to obtain non-parametrically a functional relation between y and x . However, a parametric model may also be reasonable. For example, on examining Figure 1 showing the Body Mass Index against age, we might consider the parametric model that y has a quadratic relationship to x . We may also want to test in Figure 2 if the two lines are parallel, or equivalently that the difference between the Body Mass Index and the Desired Body Mass Index for females is constant over all ages. This would involve modeling the trend lines as second degree polynomials and testing for equality in the trend lines of the parameters associated with the quadratic term as well as the parameters associated with the linear term. In all cases, the question arises as to whether or not the data can be adequately modeled by a polynomial relationship between y and x . One method that we propose as an answer to this question is to calculate the confidence

bands based on local polynomial regression. These bands can be thought of as providing a region of acceptable model representations. If an appropriate parametric regression line falls within the bands, then it provides a reasonable model description of the data. The $100(1 - \alpha)\%$ local polynomial regression bands are obtained by plotting

$$\hat{m}(x) \pm z_{\alpha/2} \sqrt{\hat{V}_p(\hat{m}(x))} \quad (6)$$

over a range of values of x , where $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, where $\hat{m}(x)$ is determined from (3) and where $\hat{V}_p(\hat{m}(x))$ is (5) with \mathbf{V} replaced by its sample estimate $\hat{\mathbf{V}}$.

The parametric regression line to be tested may be obtained in one of two ways depending upon what sample information is available. If the complete sample file with sampling weights is available, then the standard regression approach in, for example, SUDAAN may be used. If only the binned data are available, in particular the survey estimates \hat{y}_i with estimated variance-covariance matrix $\hat{\mathbf{V}}$, then another approach is needed.

For this second approach assume that $m(x_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i^T = (1, x_i, x_i^2, \dots, x_i^q)$ and where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_q)$ is the vector of regression coefficients. For the finite population we assume that $\bar{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, where the errors are deviations of the actual finite from the model. For simplicity, we assume that these errors have mean 0 and variance-covariance matrix $\sigma^2 \mathbf{I}$. Since the data are given by the survey estimates \hat{y}_i with variance-covariance matrix \mathbf{V} , the operative model is

$$\hat{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \delta_i, \quad (7)$$

where the δ_i have mean 0 and variance-covariance matrix $\sum = \sigma^2 \mathbf{I} + \mathbf{V}$. The usual weighted least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \sum^{-1} \mathbf{X})^{-1} \mathbf{X}^T \sum^{-1} \hat{\mathbf{y}}, \quad (8)$$

where the i^{th} row of \mathbf{X} is \mathbf{x}_i^T , $i = 1, \dots, k$. In terms of data analysis it is necessary to replace \sum in (8) by its estimate $\hat{\sum}$. Now the survey estimate of \mathbf{V} is $\hat{\mathbf{V}}$ so that it remains to find an estimate of σ^2 . This may be obtained through $\text{rss} = (\hat{\mathbf{y}} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\hat{\mathbf{y}} - \mathbf{X} \hat{\boldsymbol{\beta}})$, the residual sum of squares, by one of two ways.

The first method is to approximate the expected residual sum of squares under model (7) and solve directly for σ^2 . Upon using the expansion $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \dots$ we find

$$E(\text{rss}) \approx (n - q - 1) \sigma^2 + \text{tr}(\mathbf{V}) - \text{tr}(\mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}). \quad (9)$$

The estimate of σ^2 is obtained on setting rss equal to the right hand side of (8) with \mathbf{V} replaced by $\hat{\mathbf{V}}$ and then solving for σ^2 . This leads to an iterative approach to model fitting. An initial estimate of $\boldsymbol{\beta}$ is obtained from (8) with \mathbf{V} replaced by the survey estimate $\hat{\mathbf{V}}$. Then σ^2 is estimated through (9) and a new estimate of $\boldsymbol{\beta}$ using $\hat{\sum} = \hat{\sigma}^2 \mathbf{I} + \hat{\mathbf{V}}$ is obtained. The process is repeated until convergence is

obtained in the estimate of σ^2 . If the estimate of σ^2 is negative, it is set to 0. The second method for estimating σ^2 is obtaining by first treating the errors in (7) as multivariate normal variables. Then a profile likelihood for σ^2 can be obtained on replacing β and V by their estimates. The most influential term in this profile likelihood is

$$\mathbf{r}^T (\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \mathbf{r}, \quad (10)$$

where $\mathbf{r} = \hat{\mathbf{y}} - \mathbf{X}(\mathbf{X}^T(\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \mathbf{X})^{-1} \mathbf{X}^T(\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \hat{\mathbf{y}}$ is the vector of residuals. An approximation to the profile likelihood estimate $\hat{\sigma}^2$ is that value of σ^2 which minimizes (10).

To provide examples of the question of the adequacy of parametric regression, we examined two different variables in the Ontario Health Survey and their relationship to the body mass index (BMI). These were age and fat consumption as a percentage of total energy consumption. For age the binning was natural and at the precision of the recorded data. Age was restricted to the range of 18 to 65 years since the index is not applicable outside this range and age was recorded in years. The scatterplot of BMI against age with the accompanying local polynomial regression line is shown in Figure 1. The survey data on fat consumption in percentages were recorded to three decimal places. Due to the sparseness of the data at the extremes we looked at fat consumption in the range of 14 to 56% of total energy consumption. Further, we binned the data on the covariate (fat consumption) using bins 14.0 up to 14.2, 14.2 up to 14.4 and so on; the midpoints of the bins (14.1, 14.3 and so on) were used as the x_j . At each bin the survey estimate \hat{y}_i for BMI was calculated. It is the binned data that appear as a scatterplot of BMI against fat consumption in Figure 5. The solid line in Figure 5 is the local polynomial regression line with $q = 1$ for BMI on fat content. As in Figure 3, the larger variability at the extremes reflects greater sampling variability due to smaller sample sizes at the extremes. From Figure 5 it appears that BMI increases slightly as fat consumption increases. Since the complete data file for the survey was available, regression lines for all variables were obtained through SUDAAN.

In Figure 4 the solid lines are the 95% confidence bands based on (6) and the dashed line is the parametric second degree polynomial regression line. Since the dashed line falls near the border for women in their thirties and outside the bands for women in their early sixties, a second degree polynomial barely adequately describes the relation between BMI and age. Another model might be preferable. Figure 6 shows the same 95% confidence bands but for the consumption of fat as a percentage of total energy consumption. In this case the dotted line is the simple linear regression line of BMI on fat consumption. For fat consumption the line falls completely within the confidence bands so that simple linear regression appears to be an adequate description of the model relationship.

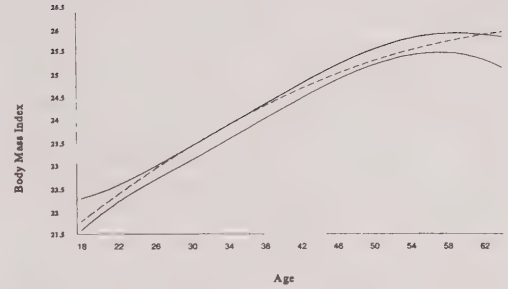


Figure 4. Confidence Bands for the Age Trend in BMI for Females

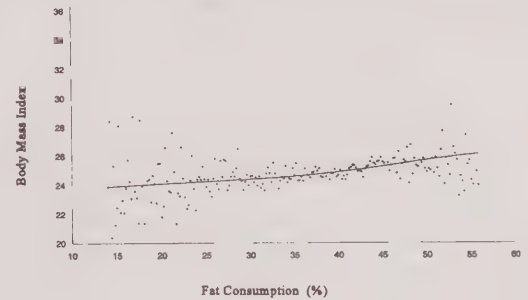


Figure 5. BMI Trend in Fat Consumption

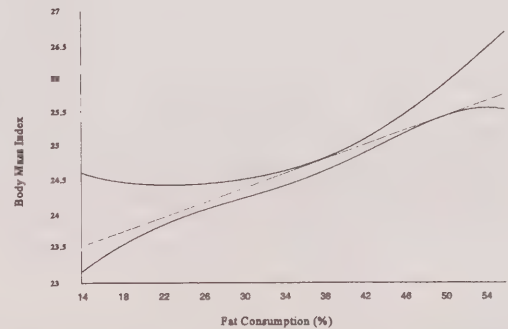


Figure 6. Confidence Bands for Fat Consumption Trend in BMI

If the data have been binned to the precision of the data as in the case of age above, and if the exploratory analysis is complete, we can stop. The estimates and variance estimates obtained are equal to the estimates and variance estimates obtained from the raw data. This may be seen on examining (3). The term on the right hand side of (3) can be expressed as a sum over the sample of the sample weights times a new measurement obtained from the raw y -measurement times an appropriate value taken from $\mathbf{e}^T(\mathbf{X}_x^T \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x$ times the total of the sample weights, where \mathbf{W}_x is \mathbf{W}_x with the p_i 's removed. These

adjusted y -measurements may be fed into SUDAAN or STATA to obtain the required approximate variance estimate. It may be that the binning has been rougher than the precision of the data or that some bins have been dropped in the tails of the distribution of x due to sparseness of the data in those bins. Both of these situations occurred in analyzing the relationship of BMI to fat consumption. Once the exploratory analysis has been completed we can return with a final model and smoothing parameter, if a nonparametric approach is used in the final analysis, and apply to model to the raw data obtaining variance estimates through SUDAAN or STATA as necessary. Depending on the amount of roughness in the binning and the number of bins dropped due to sparseness in the data, the variance estimates obtained from the raw will be approximately the same as those from the binned data.

5. FUTURE DIRECTIONS

Like Bellhouse and Stafford (1999), this paper adapts a modern method of smoothing for the analysis of complex survey data. It represents an example of a host of regression techniques that could be used. To describe these we embed the current context in a general framework hinting at future work. In doing so we mimic the developments of Hastie and Tibshirani (1990).

Here a smoother is said to be linear if fitted values are obtained by applying a matrix \mathbf{S} to a response vector \mathbf{y} . As in the case of simple linear regression for independent and identically distributed data, we let $\mathbf{H} = (\mathbf{X}^T \sum^{-1} \mathbf{X})^{-1} \mathbf{X}^T \sum^{-1}$ and further denote $(\mathbf{X}_x^T \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \hat{\mathbf{W}}_x$ as \mathbf{S}_p . Both are examples of \mathbf{S} . In addition, the response vector of binned means is a type of smooth $\hat{\mathbf{y}}_i = \mathbf{S}_b \mathbf{y}$, where \mathbf{y} is the vector of all sample responses and where \mathbf{S}_b involves the sample weights. Also the usual regression context involves applying a matrix similar to \mathbf{H} to the full response vector $\hat{\mathbf{y}}_i = \mathbf{H}_f \mathbf{y}$. So moving from usual regression to regressing means to local polynomial smoothing reduces to applying different smoothing matrices to \mathbf{y} :

$$\mathbf{H}_f \mathbf{y} \rightarrow \mathbf{H} \mathbf{S}_b \mathbf{y} \rightarrow \mathbf{S}_p \mathbf{S}_b \mathbf{y}.$$

In general \mathbf{S}_p can be replaced by any smoother \mathbf{S} and the methods extended to multiple covariates.

There are many advantages to binning the response from both a theoretical and practical standpoint. Standard smoothing tools, like those found in *Splus*, can be applied without modification of the smoother due to sampling issues. In addition, in the case of the additive model, finite population central limit theorems can be invoked and issues like degrees of freedom, choice of smoothing parameter, optimizing a criterion, can be handled in the usual manner. In the case of multiple covariates x_1, \dots, x_q the curse of dimensionality will result in sparse bins not allowing the use of the central limit theorem. This may be countered in the usual way by binning partial residuals one dimension at

a time. Here smoothers $\mathbf{S}_j \mathbf{S}_{b_j}$, $j = 1, \dots, q$ would be used in a backfitting algorithm.

It is our intention to study additive and generalized additive models in the above manner and to introduce these techniques to the analysis of complex survey data.

ACKNOWLEDGEMENTS

The authors would like to thank Rob Tibshirani for his helpful comments on this paper and the referees for their comments that helped to improve the presentation of the paper as well as to clarify some technical issues. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- BELLHOUSE, D.R., and RAO, J.N.K. (2000). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, to appear.
- BELLHOUSE, D.R., and STAFFORD, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.
- BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. Submitted for publication.
- BUSKIRK, T. (1999). *Using Nonparametric Methods for Density Estimation with Complex Survey Data*. Ph.D. dissertation, Arizona State University.
- EUBANK, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.
- FAN, J., and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.
- FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhya C*, 37, 117-132.
- GREEN, P.J., and SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.
- HARDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press: Cambridge.
- HARTLEY, H.O., and RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.
- HARTLEY, H.O., and RAO, J.N.K. (1969). A new estimation theory for sample surveys, II. In *New Developments in Survey Sampling*, N.L. Johnson and H. Smith (Eds.) New York: Wiley Inter-Science, 147-169.
- HASTIE, T.J., and TIBSHIRANI, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- KORN, E.L., and GRAUBARD, B.I. (1998). Scatterplots with survey data. *American Statistician*, 52, 58-69.
- ONTARIO MINISTRY OF HEALTH (1992). *Ontario Health Survey: User's Guide, Volumes I and II*. Queen's Printer for Ontario.

- RAO, J.N.K., and SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- ROBERTS, G., RAO, J.N.K. and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SIMONOFF, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.
- SKINNER, C.J., HOLT, D. and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.
- SMITH, T.M.F., and NJENGA, E. (1992). Robust model-based methods for analytical surveys. *Survey Methodology*, 18, 187-208.
- WAND, M.P., and JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Modelling Compositional Time Series from Repeated Surveys

D.B.N. SILVA and T.M.F. SMITH¹

ABSTRACT

A compositional time series is defined as a multivariate time series in which each of the series has values bounded between zero and one and the sum of the series equals one at each time point. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response but interest lies in the proportion of units classified in each of its categories. In this case, the survey estimates are proportions of a whole subject to a unity-sum constraint. In this paper we employ a state space approach for modelling compositional time series from repeated surveys taking into account the sampling errors. The additive logistic transformation is used in order to guarantee predictions and signal estimates bounded between zero and one which satisfy the unity-sum constraint. The method is applied to compositional data from the Brazilian Labour Force Survey. Estimates of the vector of proportions and the unemployment rate are obtained. In addition, the structural components of the signal vector, such as the seasonals and the trends, are produced.

KEY WORDS: Additive logistic transformation; Compositional time series; Kalman Filter; Labour force survey; Repeated surveys; State space models.

1. INTRODUCTION

All surveys are multivariate and multipurpose, and most are longitudinal, repeating the same questions over time. There are two broad classes of repeated surveys, those with overlapping first stage units and those with no overlap of first stage units. Both designs admit a longitudinal macro-analysis of population aggregates but only the former allows a micro-analysis and the estimation of gross flows or some other similar unit level dynamic process. In this paper we explore the time series analysis of a multivariate vector of population aggregates, a macro-analysis, while taking into account the influence of the sampling errors of the survey using disaggregated data.

Denote by $\theta_t = (\theta_{1,t}, \dots, \theta_{M+1,t})'$ a vector of population quantities of interest at time t , and assume that observations are made at equally spaced time intervals $t = 1, 2, \dots, T$. Let $y_t = (y_{1,t}, \dots, y_{M+1,t})'$ represent a survey-based estimate of θ_t , based on data collected at time t . Repeated surveys produce time series $\{y_t\}$ comprising estimates of the unknown target series $\{\theta_t\}$. Focussing on the unknown population vector θ_t , it is natural to imagine that knowledge of $\theta_1, \dots, \theta_{t-1}$ conveys useful information about θ_t but without implying that it is perfectly predictable from $\theta_1, \dots, \theta_{t-1}$. One way of representing this situation is by considering θ_t to be a random variable which evolves stochastically in time following a certain time series model, as first proposed for univariate survey analysis by Blight and Scott (1973), Scott and Smith (1974) and Scott, Smith and Jones (1977). The survey estimates y_t of θ_t can then be written as:

$$y_t = \theta_t + e_t \quad (1)$$

where $\{\theta_t\}$, $\{y_t\}$ and $\{e_t\}$ are random processes and $e_t = (e_{1,t}, \dots, e_{M+1,t})'$ are the sampling errors such that $E(e_t | \theta_t) = 0$ and $V(e_t | \theta_t) = \Sigma_t$.

The early work of Scott *et al.* (1977) was concerned with univariate $\{y_t\}$ and distinguished different forms for the data available on $\{e_t\}$. If the only data available to the analyst are the population aggregate estimates $\{y_t\}$ then this is termed a secondary analysis and the examples in Scott *et al.* (1977) are based on a secondary analysis of survey data. If the individual data records are available, then variances and covariances can be estimated directly from the data and this is called a primary analysis. In addition, in the case of a rotating panel survey, elementary estimates (based on data from a set of units that join and leave the survey at the same time) can be used to estimate the covariance structure of the sampling errors. Subsequent work by Jones (1980) used a primary analysis to measure the structure of the sampling noise whereas Binder and Hidirolou (1988), Binder and Dick (1989), Pfeffermann, Burck and Ben-Tuvia (1989), Pfeffermann and Burck (1990), Pfeffermann (1991), Binder, Bleuer and Dick (1993), Pfeffermann and Bleuer (1993), Pfeffermann, Bell and Signorelli (1996), Pfeffermann, Feder and Signorelli (1998) and Harvey and Chung (2000) employed an elementary analysis.

The time series analysis of survey data also requires that the signal process be modelled. In the early works it was assumed that $\{\theta_t\}$ was a stationary process and that $\{y_t\}$ was the superposition of two stationary processes therefore being itself stationary. Typically ARMA processes were assumed for $\{\theta_t\}$ and $\{e_t\}$, and hence for $\{y_t\}$. Binder and Hidirolou (1988) wrote the processes in state space

¹ D.B.N. Silva, Instituto Brasileiro de Geografia e Estatística, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti 106 - Rio de Janeiro, RJ Brazil, 20231-050, e-mail: denisesilva@ibge.gov.br; T.M.F. Smith, University of Southampton, Faculty of Mathematical Studies, Highfield, Southampton, SO17 1BJ, United Kingdom, e-mail: t.m.f.s@maths.soton.ac.uk.

form which led rapidly to the introduction of nonstationary processes for the signal $\{\theta_t\}$, and structural models involving trends and seasonals have been used since then.

The aim is to improve estimation of the unobservable signal and its components, but when the sampling errors are autocorrelated these autocorrelations can induce spurious trends which get confounded with the true signal trend, as pointed out by Tiller (1992) and Pfeffermann, Bell and Signorelli (1996). When the variation in the sampling errors is not taken into account, their autocorrelation structure may be absorbed into either the seasonal or the trend components, thus affecting the inference from the model.

A special case of interest in repeated surveys is when the univariate target parameter $\{\theta_t\}$ is a proportion such as the unemployment rate. Unrestricted time series modelling of $\{\theta_t\}$ may lead to estimates outside the range $0 \leq \theta_t \leq 1$. Wallis (1987) used a logistic transformation to ensure that the estimates were bounded, however he failed to take into account the survey error. Pfeffermann (1991), Tiller (1992), Pfeffermann and Bleuer (1993), Pfeffermann, Bell and Signorelli (1996) fitted state space models to unemployment rate series taking into account survey errors but without using the logistic transformation to guarantee bounded estimates.

Most surveys are multivariate and there has been little work in the multivariate time series analysis of survey data. Brunson (1987) and Brunson and Smith (1998) analyse multivariate data from opinion polls taking into account the fact that the proportions are bounded and comprise a composition, but not allowing for the structure of the survey errors. This work provides useful insight into the modelling of time series of proportions. Compositional data have also been modelled using a state space approach, by Quintana and West (1988), Shephard and Harvey (1989) and Singh and Roberts (1992), but these authors also did not address the issue of modelling the autocovariance structure of the sampling errors when the observed compositions are obtained from repeated surveys.

The motivation for this work is that many variables investigated by statistical agencies have a multinomial response and interest lies in the estimation of the proportion of units classified in each of the categories. If this is the case, the vector of proportions sums to one and forms what is known as a composition. A compositional time series is therefore a multivariate time series comprising observations of compositions at each time point. We propose a class of multivariate state space models for compositional time series from repeated surveys, which takes into account the sampling errors and guarantees estimates satisfying the underlying constraints imposed by compositions. The procedure employs a signal-plus-noise structural model which yields seasonally adjusted series and estimates of the trend which satisfy the underlying sum constraint. The method is applied to compositional data from the Brazilian Labour Force Survey comprising estimates of the vector of proportions of labour market status. Estimates of seasonally

adjusted compositions, trends and unemployment rate series are produced.

2. A FRAMEWORK FOR MODELLING COMPOSITIONAL DATA FROM OVERLAPPING SURVEYS

We assume that $\{\theta_t\}$ is multivariate and the components θ_{mt} form a composition, i.e., $0 < \theta_{mt} < 1 \forall m, t$ and $\sum_{m=1}^{M+1} \theta_{mt} = 1$. In this case y_t is a vector of sample estimates, based on the cross-sectional data of time t and belongs to the Simplex:

$$S^M = \{y_t: 0 \leq y_{mt} \leq 1, m = 1, \dots, M+1;$$

$$\sum_{m=1}^{M+1} y_{mt} = 1; t = 1, \dots, T\},$$

as in Brunson and Smith (1998). In addition, it is assumed that y_t is obtained from a survey with complex design and overlapping units between occasions. Since each of its components is subject to sampling errors, y_{mt} can be decomposed as:

$$y_{mt} = \theta_{mt} + e_{mt}, \quad m = 1, \dots, M+1, \quad (2)$$

where θ_{mt} is the unknown population proportion assumed to follow a time series model, and e_{mt} is the sampling error. Considering the $M+1$ series simultaneously, (2) can be written in vector form as in equation 1. In addition, it is assumed that

$$\sum_{m=1}^{M+1} \theta_{mt} = \sum_{m=1}^{M+1} y_{mt} = 1 \quad \forall t, \quad (3)$$

which implies that $\sum_{m=1}^{M+1} e_{mt} = 0, \forall t$.

A compositional time series is a sequence of vectors $y_t = (y_{1t}, \dots, y_{M+1,t})'$ each belonging to S^M . Aitchison (1986) examined the difficulties of applying standard methods to modelling and analysing compositions and suggested the use of transformations to map compositions from the Simplex S^M onto \mathbb{R}^M . One such transformation is the *additive logratio transformation* (a_M), defined in Aitchison (1986, page 113), which was first adopted in a time series context by Brunson (1987, page 75). The transformation is given by $v_t = a_M(y_t) = (v_{1t}, \dots, v_{Mt})'$, with

$$v_{mt} = \log \left(\frac{y_{mt}}{y_{M+1,t}} \right), \quad m = 1, \dots, M, \quad \forall t, \quad (4)$$

where \log denotes the natural logarithm. Note that $y_{M+1,t} = 1 - \sum_{m=1}^M y_{mt}$, sometimes called the fill-up value, is used as the reference variable or category. The inverse transformation, known as the *additive logistic transformation*, is given by $y_t = a_M^{-1}(v_t) = (y_{1t}, \dots, y_{M+1,t})'$ such that

$$y_{mt} = \begin{cases} \frac{\exp(v_{mt})}{1 + \sum_{j=1}^M \exp(v_{jt})} & m = 1, \dots, M, \quad \forall t, \\ \frac{1}{1 + \sum_{j=1}^M \exp(v_{jt})} & m = M+1, \quad \forall t. \end{cases} \quad (5)$$

The state space modelling procedure for compositional time series is invariant to the choice of the reference variable (Silva 1996), and so any element $y_{mt} \neq y_{M+1,t}$ of \mathbf{y}_t can be taken as the reference variable when applying the additive logistic transformation to the vector of survey estimates. When the logratios \mathbf{v}_t are normally distributed the $M+1$ – part composition has an additive logistic normal distribution as defined in Aitchison and Shen (1980). For compositional time series, Brunsdon (1987) recommended the use of Vector ARMA models (Tiao and Box 1981) for the transformed series.

We propose a procedure that not only provides predictions and filtered estimates that are bounded between zero and one and satisfy the unity-sum constraint, but also improves the estimation of the unobservable signal and its components, taking into account the sampling error.

Following Bell and Hillmer (1990), the model in (2) can be rewritten as:

$$y_{mt} = \theta_{mt} \left(1 + \frac{e_{mt}}{\theta_{mt}} \right) = \theta_{mt} u_{mt}, \quad (6)$$

with

$$u_{mt} = \left(1 + \frac{e_{mt}}{\theta_{mt}} \right) = (1 + \tilde{u}_{mt}), \quad (7)$$

where $\tilde{u}_{mt} = e_{mt}/\theta_{mt}$ represents the relative sampling error of the estimated proportion.

Applying the additive logratio transformation defined in Aitchison (1986, page 113) to the vector \mathbf{y}_t , with components given in (2), produces a transformed vector $\mathbf{v}_t = \mathbf{a}_M(\mathbf{y}_t) = (v_{1t}, \dots, v_{Mt})'$ contained in \mathbb{R}^M . If $y_{M+1,t}$ is used as the reference variable, the transformed vector has as its m^{th} component:

$$\begin{aligned} v_{mt} &= \log \left(\frac{y_{mt}}{y_{M+1,t}} \right) = \log \left(\frac{\theta_{mt} u_{mt}}{\theta_{M+1,t} u_{M+1,t}} \right) \\ &= \log \left(\frac{\theta_{mt}}{\theta_{M+1,t}} \right) + \log \left(\frac{u_{mt}}{u_{M+1,t}} \right), \quad m = 1, \dots, M. \end{aligned} \quad (8)$$

From (8), a vector model for the transformed series can be written as:

$$\mathbf{v}_t = \boldsymbol{\theta}_t^* + \mathbf{e}_t^*, \quad (9)$$

with $\mathbf{v}_t = (v_{1t}, \dots, v_{Mt})'$, $\boldsymbol{\theta}_t^* = (\theta_{1t}^*, \dots, \theta_{Mt}^*)'$ and $\mathbf{e}_t^* = (e_{1t}^*, \dots, e_{Mt}^*)'$, where $v_{mt} = \log(y_{mt}/y_{M+1,t})$, $\theta_{mt}^* = \log(\theta_{mt}/\theta_{M+1,t})$ and $e_{mt}^* = \log(u_{mt}/u_{M+1,t})$, for $m = 1, \dots, M$. Note that model (9) has the same form as model (1).

To describe the survey data, model (9) must incorporate time series models for both $\{\boldsymbol{\theta}_t^*\}$ and $\{\mathbf{e}_t^*\}$. Hence a multivariate model for the transformed data will depend on the form of the time series models for $\{\boldsymbol{\theta}_t^*\}$ and $\{\mathbf{e}_t^*\}$.

The state space formulation for compositional data is examined in section 3, the model estimation is considered in section 4 and is illustrated using Brazilian Labour Force Survey data in section 5.

3. MODELLING THE TRANSFORMED SERIES

Our approach is based on assuming that the transformed series $\mathbf{v}_t = \mathbf{a}_M(\mathbf{y}_t)$ has the signal plus noise structure in equation 9. We propose structural time series models for $\{\boldsymbol{\theta}_t^*\}$, as in Harvey (1989), and vector ARMA models (Tiao and Box 1981) for $\{\mathbf{e}_t^*\}$.

The transformed signal process $\{\boldsymbol{\theta}_t^*\}$ is assumed to follow the multivariate basic structural model, with each of the components $\{\theta_{mt}^*\}$ following a basic structural time series model (BSM) with possibly different parameters across the series. The cross-sectional relationship between the series is accounted for by the correlation structure of the system disturbances. The model for $\{\theta_{mt}^*\}$, $m = 1, 2, \dots, M$, is then given by:

$$\begin{cases} \theta_{mt}^* = L_{mt}^* + S_{mt}^* + I_{mt}^*, \\ L_{mt}^* = L_{m,t-1}^* + R_{m,t-1}^* + \eta_{mt}^{(l)}, \\ R_{mt}^* = R_{m,t-1}^* + \eta_{mt}^{(r)}, \\ S_{mt}^* = -\sum_{j=1}^{11} S_{m,t-j}^* + \eta_{mt}^{(s)}, \end{cases} \quad (10)$$

where L_{mt}^* is the trend/level component of the signal θ_{mt}^* , R_{mt}^* is the corresponding change in the level, S_{mt}^* is the seasonal component and I_{mt}^* is an irregular component. For each component, the disturbances $\eta_{mt}^{(l)}$, $\eta_{mt}^{(r)}$, $\eta_{mt}^{(s)}$, and the irregular I_{mt}^* , are assumed to be mutually uncorrelated normal deviates with mean zero and variances $\sigma_{m_1}^2$, $\sigma_{m_r}^2$, $\sigma_{m_s}^2$, $\sigma_{m_i}^2$, respectively. That is, the $M \times 1$ vector disturbances $\boldsymbol{\eta}_t^{(l)}$, $\boldsymbol{\eta}_t^{(r)}$, $\boldsymbol{\eta}_t^{(s)}$ and \mathbf{I}_t^* are mutually uncorrelated in all time periods. In addition, the irregulars I_{mt}^* , $I_{j(t-h)}^*$, with $m \neq j$, $h = \dots, -2, -1, 0, 1, 2, \dots$, are assumed to be correlated when $h = 0$, but uncorrelated for $h \neq 0$ and \mathbf{I}_t^* has covariance matrix Σ_I . The same happens with the

system disturbances $\eta_{mt}^{(a)}$, $\eta_{j(t-h)}^{(a)}$, $a = l, r, s$, which are also correlated when $h=0$, but uncorrelated for $h \neq 0$, with covariance matrices \sum_l, \sum_r, \sum_s . At each time t , the correlation structure between the components of the composition is summarized by \sum_l and a block diagonal matrix with the blocks being \sum_l, \sum_r, \sum_s . Note that the relation between the series arises via the non-zero off-diagonal elements of the disturbance covariance matrices. The multivariate model (10) for $\{\theta_t^*\}$ has the following state space formulation:

$$\begin{cases} \theta_t^* = H^{(0)} \alpha_t^{(0)} + I_t^*; \\ \alpha_t^{(0)} = T^{(0)} \alpha_{t-1}^{(0)} + G^{(0)} \eta_t^{(0)}, \end{cases} \quad (11)$$

where $H^{(0)} = [101000000000] \otimes I_M$,

$$\alpha_t^{(0)} = [L_{1t}^* \dots L_{Mt}^* R_{1t}^* \dots R_{Mt}^* S_{1t}^* \dots S_{Mt}^* S_{1,t-10}^* \dots S_{M,t-10}^*]'$$

$$\eta_t^{(0)} = (\eta_{1t}^{(l)} \dots \eta_{Mt}^{(l)} \eta_{1t}^{(r)} \dots \eta_{Mt}^{(r)} \eta_{1t}^{(s)} \dots \eta_{Mt}^{(s)})',$$

$$G^{(0)} = \begin{bmatrix} I_3 \\ \dots \\ \mathbf{0}_{10 \times 3} \end{bmatrix} \otimes I_M,$$

$$T^{(0)} = \begin{bmatrix} 1 & 1 & : & & & \mathbf{0}_{2 \times 11} \\ 0 & 1 & : & & & \\ \dots & \dots & : & \dots & \dots & \dots \\ & & : & -1 & -1 & \dots & -1 & -1 \\ & & : & 1 & 0 & \dots & 0 & 0 \\ \mathbf{0}_{11 \times 2} & : & 0 & 1 & \dots & 0 & 0 \\ & : & : & : & : & : & : \\ & : & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \otimes I_M.$$

The transformed survey error process $\{e_t^*\}$ is assumed to follow an M -dimensional vector autoregressive moving average process (VARMA), defined by $\Phi(B)e_t^* = \Theta(B)a_t$, with mean vector $E(e_t^*) = \mathbf{0}$ and

$$\Theta(B) = I - \Theta_1 B - \dots - \Theta_q B^q,$$

$$\Phi(B) = I - \Phi_1 B - \dots - \Phi_p B^p,$$

where $\Phi_1, \dots, \Phi_p, \Theta_1, \dots, \Theta_q$ are coefficient matrices and a_t is an M -dimensional white noise random vector with zero mean and covariance structure:

$$E(a_t a_{t-h}') = \begin{cases} \sum_a & h = 0 \\ \mathbf{0} & h \neq 0 \end{cases}.$$

The cross-covariance matrix function for the VARMA process $\{e_t^*\}$, (see Wei 1993, page 333), is given by:

$$\Gamma_{e^*}(h) = \text{COV}(e_{t-h}^*, e_t^*) = E(e_{t-h}^* e_t^{*'}),$$

where $\{\Gamma_{e^*}(h)\}_{m,j} = \gamma_{e^*,m,j}(h) = \text{COV}(e_{m,t-h}^*, e_{jt}^*)$, and the cross-correlation function for the vector process is defined as:

$$P_{e^*}(h) = D_{e^*}^{-1/2} \Gamma_{e^*}(h) D_{e^*}^{-1/2},$$

where

$$D_{e^*} = \text{diag}(\gamma_{e^*,11}(0), \dots, \gamma_{e^*,MM}(0)).$$

The state space representation of VARMA models can be found in Reinsel (1993, section 7.2). The separate models for the transformed signal and sampling errors can be cast into a unique state space model, see Silva (1996, Chapter 8) for details.

4. ESTIMATION FROM THE TRANSFORMED DATA

As in previous sections, we distinguish between the estimation of the structure of the surveys errors, the noise, and the estimation of the covariances of the basic structural model. Once these are obtained, we employ the Kalman filter to get estimates of the trend and seasonals which determine the signal. Before carrying out the signal extraction, the VARMA model for the survey errors must be identified.

The model specification for the error process $\{e_t^*\}$ depends on the sampling design, particularly on the level of sample overlap between occasions, and also on data availability. Many authors have considered the problem of modelling the sampling error process in a univariate framework, see, for example, Scott and Smith (1974), Pfeiffermann (1989, 1991), Bell and Hillmer (1990), Binder and Dick (1989), Tiller (1989, 1992), Pfeiffermann and Bleuer (1993), Binder, Bleuer and Dick (1993), Pfeiffermann, Bell and Signorelli (1996) and Pfeiffermann, Feder and Signorelli (1998). However, in all of these cases the authors are working with the original data instead of the transformed data. After transformation, it is difficult to carry out a full primary analysis based on individual observations, see Silva (1996, Chapter 7).

Many repeated surveys are based on a rotating panel design in which K panels of sampling units are investigated at each survey round (time point) and panels are replaced in a systematic manner, according to the rotating pattern of the survey design. In these surveys, elementary design unbiased estimates $y_t^{(k)}$, $k = 1, \dots, K$, for the population parameter θ_p , can be obtained from each rotation group. A rotation group is a set of sampling units that joins and leaves the sample at the same time.

In a two-stage survey, in which the primary sampling units (enumeration areas) remain in the sample for all survey occasions, the replacement of panels of households (second-stage units) is ordinarily carried out within geographical regions defined by mutually exclusive groups of enumeration areas. Note that a survey with K panels produces K streams of estimates, where a stream is a time series of all sample estimates based on samples from the same enumeration area, that is, is a time series of elementary estimates.

Pfeffermann, Bell and Signorelli (1996) and Pfeffermann, Feder and Signorelli (1998) show how to estimate the autocorrelation of the sampling error process for univariate data, before transformation, using the so-called pseudo-errors, defined as:

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t, \quad (12)$$

where $y_t = 1/K \sum_{k=1}^K y_t^{(k)}$. If there is no rotation bias, it follows that:

$$\tilde{e}_t^{(k)} = e_t^{(k)} - e_t, \quad (13)$$

thus contrasts in $y_t^{(k)}$ are contrasts in the panel sampling errors $e_t^{(k)}$.

For the compositional case we apply, for each elementary estimate, the transformation $v_t^{(k)} = a_m(y_t^{(k)}) = (v_{1t}^{(k)}, \dots, v_{Mt}^{(k)})'$ which has as its m^{th} component, ($m = 1, \dots, M$):

$$v_{mt}^{(k)} = \log \left(\frac{y_{mt}^{(k)}}{y_{M+1,t}^{(k)}} \right) = \log \left(\frac{\theta_{mt}}{\theta_{M+1,t}} \right) + \log \left(\frac{u_{mt}^{(k)}}{u_{M+1,t}^{(k)}} \right). \quad (14)$$

From (14), a vector model for the k^{th} series of transformed elementary estimates can be written as:

$$v_t^{(k)} = \theta_t^* + e_t^{*(k)}, \quad (15)$$

with $e_t^{*(k)} = (e_{1t}^{*(k)}, \dots, e_{Mt}^{*(k)})'$ and $e_{mt}^{*(k)} = \log(u_{mt}^{(k)} / u_{M+1,t}^{(k)})$, for ($m = 1, \dots, M$). Hence, from (15), M -dimensional time series of transformed pseudo-errors can be constructed from deviations of the transformed rotation group estimates about their overall mean. The transformed pseudo-errors for the k^{th} rotation group are defined as:

$$\begin{aligned} \tilde{e}_t^{*(k)} &= (\tilde{e}_{1t}^{*(k)}, \dots, \tilde{e}_{Mt}^{*(k)})' = v_t^{(k)} - v_t \\ &= (v_{1t}^{(k)} - v_{1t}, \dots, v_{Mt}^{(k)} - v_{Mt})', \end{aligned} \quad (16)$$

where $v_t = 1/K \sum_{k=1}^K v_t^{(k)}$. Note, in addition, that $\tilde{e}_t^{*(k)} = e_t^{(k)} - e_t$.

From (14) and (15), it becomes clear that the framework introduced by Pfeffermann, Bell and Signorelli (1996) can also be applied to the transformed model.

The cross-correlation matrices of the transformed sampling errors can be obtained by averaging the cross-

covariances matrices of the transformed pseudo-errors as follows (for details see Silva 1996, Chapter 7):

$$P_{e^*}(h) = \left[\sum_{k=1}^K D_{\tilde{e}^*}^{(k)} \right]^{-1/2} \left[\sum_{k=1}^K \Gamma_{\tilde{e}^*}^{(k)}(h) \right] \left[\sum_{k=1}^K D_{\tilde{e}^*}^{(k)} \right]^{-1/2}, \quad (17)$$

where

$$\Gamma_{\tilde{e}^*}^{(k)}(h) = \text{COV}(\tilde{e}_{t-h}^{*(k)}, \tilde{e}_t^{*(k)}) = E(\tilde{e}_{t-h}^{*(k)} \tilde{e}_t^{*(k)'}),$$

with

$$\{\Gamma_{\tilde{e}^*}^{(k)}(h)\}_{mj} = \text{COV}(\tilde{e}_{m,t-h}^{*(k)}, \tilde{e}_{jt}^{*(k)}) = \gamma_{\tilde{e}^*,mj}^{(k)}(h)$$

and

$$D_{\tilde{e}^*} = \text{diag}(\gamma_{\tilde{e}^*,11}^{(k)}(0), \dots, \gamma_{\tilde{e}^*,MM}^{(k)}(0)).$$

Once the correlation matrices $P_{e^*}(h)$, $h = 1, 2, \dots$ have been estimated, a VARMA model to represent the transformed survey error process can be selected and estimates of the respective parameter matrices can be computed, provided the series of transformed pseudo-errors are available. Then, as described in section 3, a state space model for representing the transformed signal and sampling errors can be defined and the Kalman filter equations can be used to get filtered and smoothed estimates for the unobservable components. The application of the Kalman Filter requires the estimation of the unknown hyperparameters (the covariance matrices $\sum_t, \sum_r, \sum_s, \sum_l, \sum_a$) and the estimation of the initial state vector and the respective covariance matrices.

Having addressed the issue of how to model the survey estimates in a compositional framework and how to identify the time series model for the transformed sampling errors, the following section presents the results of an empirical study using compositional data from the Brazilian Labour Force Survey.

5. MODELLING COMPOSITIONAL TIME SERIES IN THE BRAZILIAN LABOUR FORCE SURVEY

The Brazilian Labour Force Survey (BLFS) collects monthly information about employment, hours of work, education and wages together with some demographic information. It classifies the survey respondents, aged 15 and over, according to their employment status in the week prior to the interview into three main groups: employed, unemployed and not in the labour force, following the International Labour Organization (ILO) definitions. The survey targets the population living at the six major metropolitan areas in the country. The BLFS is a two-stage sample survey in which the primary sampling units (psu) are the census enumeration areas (EA) and the second-stage units (ssu) are the households. The primary sampling units

are selected with probabilities proportional to their sizes and then a fixed number of households is selected from each sampled EA by systematic sampling. All household members within the selected households are enumerated. The primary sampling units remain the same for a period of roughly 10 years (as in a master sample). New primary sampling units are selected when information from a new population census becomes available.

In addition, the BLFS is a rotating panel survey. For any given month the sample is composed of four rotation groups of mutually exclusive sets of primary sampling units. The rotation pattern applies to panels of second-stage units (households). Within each rotation group a panel of households stays in the sample for four successive months, is rotated out for the following 8 months and then is sampled again for another spell of four successive months. Each month one panel is rotated out of the sample. The substituting panel can be a new panel or one that has already been observed for the first four months period. Note that the 4-8-4 rotation pattern induces a complex correlation structure for the sampling errors over time and that there is a 75% overlap between two successive months.

The empirical work was carried out using data from the São Paulo metropolitan area covering the period from January 1989 to September 1993 (57 observations). The quantities of interest are the proportions of employed, unemployed and not in the labour force, and also the unemployment rate. Using the monthly individual observations, the series of sample estimates and their respective estimated standard errors were computed using data of each specific survey round and standard estimators. For each month, two sets of estimates were obtained. The direct sample

estimates, derived from the complete data collected at a given month and four elementary estimates, each based on data from a single rotation group. The panel estimates are used to estimate the sampling error autocorrelations and to help to identify the time series model for the sampling errors.

In this study the observed composition has $M + 1 = 3$ components and the time series is defined as the sequence of vectors $y_t = (y_{1t}, y_{2t}, y_{3t})'$, where:

y_{1t} is the estimated proportion of unemployed persons in month t ;

y_{2t} is the estimated proportion of employed persons in month t ;

y_{3t} is the estimated proportion of persons not in the labour force in month t .

The model for the BLFS must incorporate the special features of the data. Firstly, it is a compositional time series belonging to the Simplex S^2 at each time t . Secondly, the time series are subject to sampling errors. Following the approach in section 2, we first map the composition onto \mathbb{R}^2 using the additive logratio transformation with y_{3t} as the reference category. As y_t is a vector of sample estimates, it can be modelled as in equation 1 and the vector model for the transformed series is given by equation 9. Then, the transformed composition is modelled using a multivariate state space model that accounts for the autocorrelations between the sampling errors. Finally, the model based estimates are transformed back to the original space. Figure 1 displays the series of transformed compositions.

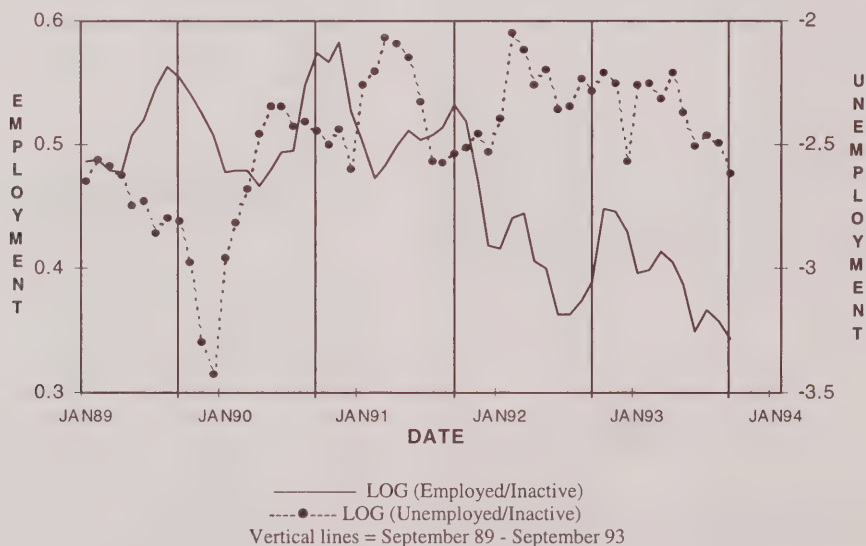


Figure 1. Brazilian Labour Force Series - SÃO PAULO Transformed Compositions

The model for the transformed sample estimates \mathbf{v}_t is composed of a bivariate model for the transformed signal $\boldsymbol{\theta}_t^*$, describing how the transformed population quantities evolve in time, and a bivariate model representing the time series relationship between transformed sampling errors \mathbf{e}_t^* . The transformed signal process $\{\boldsymbol{\theta}_t^*\}$ is assumed to follow the bivariate basic structural model (equation 11) as described in section 3. As mentioned before, a VARMA model to represent the sampling error series was used. The correlation structure of the transformed sampling errors was estimated using the transformed pseudo-errors as in equation 16. In addition, estimates of the partial lag correlation matrices for $\{\mathbf{e}_t^*\}$ were computed using a recursive algorithm provided in Wei (1993, pages 359-362). A program in SAS-IML which gives the corresponding schematic representations (Tiao and Box 1981) and a statistical test to help establish the order of the vector process was developed. The form of the correlation matrices and the results for the statistical test, available in Silva (1996), indicate that a VAR(1), a VAR(2) or a VARMA(1,1) model could be used to represent the transformed sampling error process. In the event, the VARMA(1,1) was chosen because it yields smaller standard errors for estimates of the unemployment rate. The parameter estimates for this model were obtained from the relationship between the cross-covariance function and the parameter matrices given in Wei (1993, pages 346-347). The VARMA(1,1) fitted for $\{\mathbf{e}_t^*\}$ is given by:

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} 0.7347 & 0.2414 \\ -0.9224 & -0.2072 \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} - \begin{bmatrix} 0.3162 & 0.2590 \\ -0.7666 & -0.2749 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix},$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001723 & 0.0003476 \\ 0.0003476 & 0.0051660 \end{bmatrix} \tag{18}$$

Having put the combined model for the transformed survey estimates into the state space form, the Kalman Filter equations can be used to get filtered and smoothed estimates for the unobservable components. Note that the estimation of the model for the transformed sampling errors (equation 18) was implemented outside the Kalman Filter. The application of the Kalman Filter requires the estimation of the unknown hyperparameters (the covariances), the initial state vector and respective covariance matrix. Assuming that the disturbances $\boldsymbol{\eta}_t^{(0)}$, \mathbf{a}_t , and \mathbf{I}_t are normally distributed, the log-likelihood function of the (transformed) observations can be expressed via the prediction error

decomposition (for details see Harvey 1989). Estimates for the model covariances were obtained by maximum likelihood, applying a quasi-Newton optimization technique. A computer program to implement the maximization procedure was developed using the optimization routine NLPQN from SAS-IML.

The initialization of the Kalman filter was carried out using a combination of a diffuse and proper priors. Following this approach, the non-stationary components $(\boldsymbol{\alpha}^{(0)})'$ of the state vector were initialized with very large error variances and the respective components of the initial state vector were taken as zero. The stationary components $(e_{1t}^*, e_{2t}^*)'$ were initialized by the corresponding unconditional mean and variance.

When fitting the model, the estimated covariance matrices obtained for the slope and seasonal components were very small and could be set to zero. This implies that the seasonals are assumed to be deterministic and that the slope is assumed to be fixed, giving rise to a local level model with a drift and non-stochastic seasonals for the signal. Indeed, as pointed out by Koopman, Harvey, Doornik and Shephard (1995, page 39), when the number of years considered in the analysis is small, it seems reasonable to fix the seasonals since there is not enough data to allow the estimation of a changing pattern. The fact that a fixed seasonal pattern is validated by the estimation process is a satisfactory feature of the modelling procedure. In addition, the estimated covariance matrix of the irregular component was also found to be very small (and hence undetectable) in comparison to the sampling error and so, as expected, in the presence of relatively large sampling errors, there was no need to include irregular components in the model for the transformed signal. The parameter estimates and respective asymptotic errors (displayed in parenthesis) are presented in Table 1.

Table 1
Estimates for the Hyperparameters and Standard Errors

Model	$\hat{\Sigma}_I \times 10^{-4}$ (2)	$\hat{\Sigma}_r = \hat{\Sigma}_s = \hat{\Sigma}_I$
BSM + VARMA (1,1)	$\begin{bmatrix} 2.78 & \mathbf{0.12} \\ (0.91) & \\ 1.95 & 87.0 \\ (3.55) & (27.10) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$
(1)		

- (1) Local level model with drift and fixed seasonals for the signal.
- (2) Upper-triangular contains correlation.

To evaluate the model performance, empirical distributions of the standardized residuals were compared with a standard normal distribution to verify the assumption that the innovations $(\mathbf{v}_t - \hat{\mathbf{v}}_{t|t-1})$ are normal deviates. Examination of corresponding normal plots revealed no departure from normality. In addition, we also computed the auto-

correlations of the innovations, which were close to zero, further validating the model.

Predictions for $y_{m,t}$ and estimates for $\theta_{m,t}$ are computed by applying the additive logistic transformation (equation 5) to predictions $\hat{v}_{t|t-1}$ and smoothed estimates $\hat{\theta}_{t|T}^*$ for the transformed series and signal, respectively. This transformation maps these estimates onto S^2 , guaranteeing that they satisfy the boundedness constraints.

Unfortunately, although $\hat{L}_{t|T}^*$ and $\hat{S}_{t|T}^*$ can be obtained from $\hat{\theta}_{t|T}^*$, it is not straightforward to obtain estimates for the structural unobservable components of the original signal θ_t , such as $\hat{L}_{t|T}$ and $\hat{S}_{t|T}$. However, if a multiplicative model with no irregular component is assumed for $\{\theta_{m,t}\}$, such that:

$$\theta_{1t} = L_{1t} S_{1t}, \theta_{2t} = L_{2t} S_{2t}, \theta_{3t} = L_{3t} S_{3t}, \quad (19)$$

where $L_{m,t}$ and $S_{m,t}$, for $m = 1, 2, 3$ represent the trend and seasonal components of the unobservable signals, then applying an additive logratio transformation to θ_t results in:

$$\begin{aligned} \log(\theta_{m,t} / \theta_{3t}) &= \log\left(\frac{L_{m,t} S_{m,t}}{L_{3t} S_{3t}}\right) \\ &= \log\left(\frac{L_{m,t}}{L_{3t}}\right) + \log\left(\frac{S_{m,t}}{S_{3t}}\right), m = 1, 2. \end{aligned} \quad (20)$$

This can be rewritten as:

$$\theta_{m,t}^* = L_{m,t}^* + S_{m,t}^*, \quad (21)$$

with $L_{m,t}^* = \log(L_{m,t} / L_{3t})$ and $S_{m,t}^* = \log(S_{m,t} / S_{3t})$. Thus, the use of a basic structural model for $\{\theta_t^*\}$ corresponds to the case in which the underlying model for $\{\theta_t\}$ decomposes the original signal into its trend and seasonal components in a multiplicative way. For deriving estimates, either filtered or smoothed, for $L_{m,t}$ note that:

$$\exp(L_{1t}^*) = L_{1t} / L_{3t}, \quad \exp(L_{2t}^*) = L_{2t} / L_{3t}. \quad (22)$$

To recover L_{1t} , L_{2t} , L_{3t} , in (22), it is necessary to assume an explicit relationship between these unobservable components based on model (19). By doing this, a third equation can be added to the system in (22) and an estimate of the original series components can be obtained. Note that the system has three unknowns for just two equations. In this case, it is quite natural to assume that the level components sum to one across the series, being also bounded between zero and one. Hence, trend estimates for the original series can be obtained solving:

$$\begin{cases} \exp(L_{1t}^*) &= L_{1t} / L_{3t}, \\ \exp(L_{2t}^*) &= L_{2t} / L_{3t}, \\ L_{1t} + L_{2t} + L_{3t} &= 1, \end{cases} \quad (23a)$$

which results in

$$\begin{aligned} L_{m,t} &= \frac{\exp(L_{m,t}^*)}{1 + \sum_{k=1}^2 \exp(L_{k,t}^*)}, \quad m = 1, 2; \\ L_{3t} &= \frac{1}{1 + \sum_{k=1}^2 \exp(L_{k,t}^*)}. \end{aligned} \quad (23b)$$

As there is no irregular component in model (19) the seasonally adjusted figures are given by the trend estimates in (23). Therefore, the smoothed estimates for the trend of the original series of proportions are obtained by applying the additive logistic transformation to $\hat{L}_{t|T}^*$. Consequently, estimates for the seasonal components of the original proportions can be computed as:

$$\hat{S}_{m,t|T} = \hat{\theta}_{m,t|T} / \hat{L}_{m,t|T}, \quad m = 1, 2, 3.$$

For labour force surveys, an important issue is the estimation of the unemployment rate series (as opposed to unemployment proportions) and also the production of the corresponding seasonally adjusted figures. Recall that θ_{1t} and θ_{2t} represent the unknown population proportions of unemployed and employed people, respectively. Using these proportions, the unknown unemployment rate at time t is defined as

$$R_t = \frac{\theta_{1t}}{\theta_{1t} + \theta_{2t}} = \frac{1}{\left(1 + \frac{\theta_{2t}}{\theta_{1t}}\right)} = \left(\frac{\theta_{2t}}{\theta_{1t}} + 1\right)^{-1}. \quad (24)$$

Based on model (11), trend estimates for the unemployment rate can be obtained by simply replacing $\theta_{m,t}$ by $L_{m,t}$, $m = 1, 2$, in equation 24. In conclusion, the methodology developed in this section provides signal (and trend) estimates that are bounded between zero and one and satisfy the unit-sum constraint. It also provides estimates for the seasonal and trend components of series comprising ratios of the original proportions which is a useful feature.

Figure 2 presents the design-based estimates and the model-dependent estimates for the proportion of unemployed persons, for the time period January 1989 to September 1993. The model-dependent estimates are the smoothed estimates which use all the data for the whole sample period. As can be seen from the graph, the signal estimates behave similarly to the design-based estimates although some of the sharp turning points in the series have been smoothed out.

Model-dependent trend estimates were obtained by fitting the basic structural model defined for the signal process when sampling error variation was modelled as a VARMA(1,1). These estimates were compared with the estimates produced by the familiar X-11 procedure. Figure 3 displays the trend produced for the unemployment rate

series by both methods together with the estimates obtained by fitting a standard basic structural model which does not account for sampling error variation.

The trend produced by our model is smoother, suggesting that the model-dependent procedure succeeds in removing the fluctuations induced by the sampling errors.

In addition, model-dependent estimates for the seasonal effects of the original compositions were also obtained from the multivariate modelling procedure which accounts for two very important features of the data, namely the compositional constraints and the presence of sampling errors.

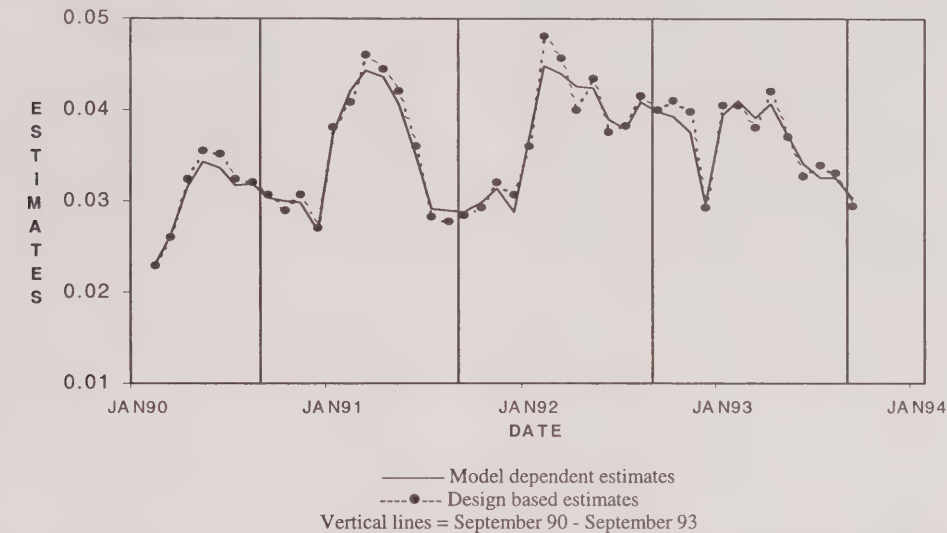


Figure 2. Brazilian Labour Force Series - SÃO PAULO Design Based and Model Dependent Estimates Proportion of Unemployed Persons

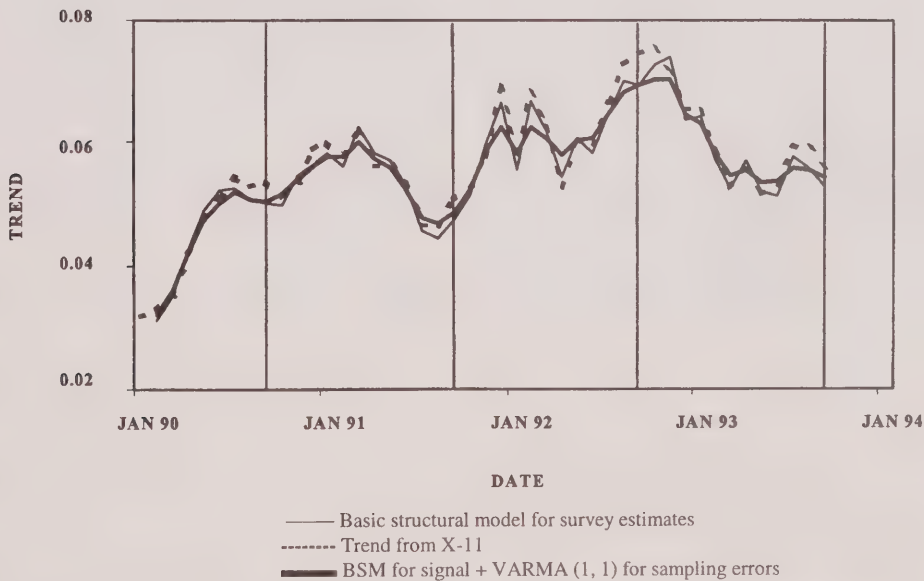


Figure 3. Brazilian Labour Force Series - SÃO PAULO Trend Estimates for the Unemployed Rates Series

6. CONCLUSIONS

This paper proposes a state space approach for modelling compositional time series from repeated surveys. The important feature of the proposed methodology is that it provides bounded predictions and signal estimates of the parameters in a composition, satisfying the unity-sum constraint, while taking into account the sampling errors. This is accomplished by mapping the compositions from the Simplex onto Real space using the additive logratio transformation, modelling the transformed data employing multivariate state space models, and then applying the additive logistic transformation to obtain estimates in the original scale.

The empirical work using data from the Brazilian Labour Force Survey demonstrates the usefulness of this modelling procedure in a genuine survey situation, showing that it is possible to model the multivariate system and obtain estimates for all the relevant components. The results of the empirical work also show that smoother trends and fixed seasonals are obtained from a model which explicitly accounts for the sampling errors, when compared with estimates produced by X-11. In addition, because the model-dependent estimators combine past and current survey data, the standard deviations of these estimates are in general lower than the standard deviations of the design-based estimators, as shown in Silva (1996, Chapter 8).

One drawback of the proposed procedure is that although confidence regions for the original compositional vector can be constructed based on the model-dependent estimates by using the additive logistic normal distribution, confidence intervals for the individual proportions are not readily available. Such intervals could be obtained from marginal distributions of the additive logistic normal distribution, but these can only be evaluated by integrating out some of the elements of the compositional vector and, as pointed out by Brunson (1987, page 135), this produces intractable expressions.

Under a state space formulation a wide variety of models is available to represent the multivariate signal and noise processes, which is a great benefit of this modelling procedure. The application of the method to different data sets is recommended. Further empirical research should also consider situations where the composition lies on a Simplex with dimensions higher than two and/or with compositions evolving close to the boundaries of the interval [0,1]. In addition, a better insight into the performance of the modelling procedure may be gained by applying the method to simulated data, for which the "true" underlying models are known. The models considered here can also be extended to incorporate rotation group bias effects and explanatory variables.

ACKNOWLEDGEMENTS

This research was supported by CAPES-Brazil and IBGE-Brazil and by a grant from the Economic and Social

Research Council of the UK under its Analysis of Large and Complex Datasets Programme. The authors wish to thank the referees and Prof. Danny Pfeffermann for the suggestions that led to many improvements in the paper. Thanks are also due to Dr. Harold Mantel for his encouragement towards the preparation of the final version.

REFERENCES

- AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. New York: Chapman and Hall.
- AITCHISON, J., and SHEN, S.M. (1980). Logistic-Normal distributions: some properties and uses. *Biometrika*, 67, 261-272.
- BELL, W.R., and HILLMER, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, 195-215.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. In *Handbook of Statistics*, (Eds., P.R. Krishnaiah and C.R. Rao). Elsevier Science, 6, 187-211.
- BINDER, D.A., and DICK, J. P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BINDER, D.A., BLEUER, S.R. and DICK, J.P. (1993). Time series methods applied to survey data. *Proceedings of the 49th International Statistical Institute Session*, 1, 327-344.
- BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, B*, 35, 61-68.
- BRUNSDON, T.M. (1987). Time Series Analysis of Compositional Data. Unpublished Ph.D. Thesis. University of Southampton.
- BRUNSDON, T.M., and SMITH, T.M.F. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14, 3, 237-253.
- DE JONG, P. (1988). The likelihood for a state space model. *Biometrika*, 75, 165-169.
- DE JONG, P. (1989). Smoothing and interpolation with the state space model. *Journal of the American Statistical Society*, 84, 1085-1088.
- DE JONG, P. (1991). The diffuse Kalman filter. *The Annals of Statistics*, 19, 1073-1083.
- FERNANDEZ, F.J.M., and HARVEY, A.C. (1990). Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business and Economic Statistics*, 8, 1, 71-81.
- GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Social Statistics Section*, 242-257.
- HARRISON, P.J., and STEVENS, C.F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, B*, 38, 205-47.
- HARVEY, A.C. (1986). Analysis and generalisation of a multivariate exponential smoothing model. *Management Science*, 32, 374-380.
- HARVEY, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press. Cambridge.

- HARVEY, A.C. (1993). *Time Series Models*. Second Edition. Harvester Wheatsheaf. London.
- HARVEY, A.C., and PETERS, S. (1984). Estimation Procedures for Structural Time Series Models. London School of Economics. Mimeo.
- HARVEY, A.C., and SHEPHARD, N. (1993). Structural time series models. In *Handbook of Statistics*, (Eds. S. Maddala, C.R. Rao and H.D. Vinod). Elsevier Science Publishers, 11, 261-302.
- HARVEY, A.C., and CHUNG, C.(2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, A*, 163, Part 3, 303-339.
- IBGE (1980). Metodologia da Pesquisa Mensal de Emprego 1980. Relatórios Metodológicos. Fundação Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro.
- JONES, R.G.(1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, B*, 42, 221-226.
- KOOPMAN, S.J., HARVEY, A.C., DOORNIK, J.A. and SHEPHARD, N. (1995). *STAMP 5.0 - Structural Time Series Analyser, Modeller and Predictor*. Chapman & Hall. London.
- MITTNIK, S. (1991). Derivation of the unconditional state covariance matrix for exact-likelihood estimation of ARMA models. *Journal of Economic Dynamics and Control*, 15, 731-740.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-177.
- PFEFFERMANN, D., BURCK, L. and BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Proceedings of the International Symposium on Analysis of Data in Time*, 43-55.
- PFEFFERMANN, D., and BURCK, L.(1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 217-237.
- PFEFFERMANN, D., and BLEUER, S.R. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 149-164.
- PFEFFERMANN, D., BELL, P. and SIGMORELLI, D. (1996). Labour force trend estimation in small areas. *Proceedings of the Annual Research Conference, Bureau of the Census*, 407- 431.
- PFEFFERMANN, D., FEDER, M. and SIGMORELLI, D. (1998). Estimation of autocorrelations of survey errors with applications to trend estimation in small samples. *Journal of Business and Economics Statistics*, 16, 339-348.
- QUINTANA, J.M., and WEST, M. (1988). Time series analysis of compositional data. *Journal of Bayesian Statistics*, (Eds. J.H. Bernardo, M.A. Degroot and A.F.M. Smith). Oxford University Press, 3.
- REINSEL, G.C. (1993). *Elements of Multivariate Time Series Analysis*. Springer-Verlag.
- SAS INSTITUTE INC. (1995). *SAS/IML Software: Changes and Enhancements through Release 6.11*. SAS institute Inc. Cary, NC.
- SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.
- SCOTT, A.J., SMITH, T.M.F. and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- SHEPHARD, N.G., and HARVEY, A.C. (1989). Tracking the Level of Support for the Parties During General Election Campaigns. Mimeo. Dept. of Statistics, London School of Economics.
- SILVA, D.B.N. (1996). Modelling Compositional Time Series From Repeated Surveys. Unpublished PhD Thesis. University of Southampton. UK.
- SINGH, A.C., and ROBERTS, G.R. (1992). State space modelling of cross-classified time series of counts. *International Statistical Review*, 60, 321-335.
- SMITH, T.M.F., and BRUNSDON, T.M. (1986). Time Series Methods for Small Areas. Unpublished Report. University of Southampton.
- TIAO, G.C., and BOX, G.E.P. (1981). Modelling multiple time series with applications. *Journal of the American Statistical Association*, 76, 802-816.
- TILLER, R.B. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 16-25.
- TILLER, R.B.(1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 2, 149-166.
- WALLIS, F. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, 8, 115-23.
- WEI, W.W.S. (1993). *Time Series Analysis - univariate and multivariate methods*. Addison-Wesley.
- WEST, M., and HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 2001. An asterisk indicates that the person served more than once.

- | | |
|--|---|
| M. Axelson, <i>Örebro University, Sweden</i> | * P. Lavallée, <i>Statistics Canada</i> |
| M. Bankier, <i>Statistics Canada</i> | H. Lee, <i>Westat, Inc.</i> |
| K. Brewer, <i>Australian National University</i> | J. Lent, <i>U.S. Bureau of Transportation Statistics</i> |
| Moon Jung Cho, <i>U.S. Bureau of Labor Statistics</i> | * S. Lohr, <i>Arizona State University</i> |
| R. Chambers, <i>University of Southampton</i> | W. Lu, <i>Simon Fraser University</i> |
| * S. Chowdhury, <i>Westat, Inc.</i> | J. Moloney, <i>Statistics Canada</i> |
| M. P. Cohen, <i>U.S. Bureau of Transportation Statistics</i> | G. Montanari, <i>University of Perugia</i> |
| G. Datta, <i>University of Georgia</i> | C. Perry, <i>NASS</i> |
| P. Duchesne, <i>École des Hautes Études Commerciales de Montréal</i> | T.E. Raghunathan, <i>University of Michigan</i> |
| F. Dupont, <i>INSEE</i> | E. Rancourt, <i>Statistics Canada</i> |
| M.R. Elliott, <i>University of Pennsylvania</i> | L.-P. Rivest, <i>Université Laval</i> |
| J.L. Eltinge, <i>U.S. Bureau of Labor Statistics</i> | N. Schenker, <i>National Center for Health Statistics</i> |
| * V. M. Estevao, <i>Statistics Canada</i> | J. Sedransk, <i>Case Western University</i> |
| M. Ghosh, <i>University of Florida</i> | * A.C. Singh, <i>Research Triangle Institute</i> |
| B. Graubard, <i>National Cancer Institute</i> | K.P. Srinath, <i>ABT Associates</i> |
| R. Harter, <i>National Opinion Research Center</i> | B. Sutrahur, <i>Memorial Univesity</i> |
| M.A. Hidiroglou, <i>Statistics Canada</i> | * A. Théberge, <i>Statistics Canada</i> |
| B. Hulliger, <i>Swiss Federal Statistical Office</i> | R. Thomas, <i>Carleton University</i> |
| D. Jang, <i>Mathematica Policy Research</i> | S. K. Thompson, <i>Pennsylvania State University</i> |
| D. Kostanich, <i>U.S. Bureau of the Census</i> | C. Tucker, <i>U.S. Bureau of Labor Statistics</i> |
| P. Kott, <i>NASS</i> | * R. Valliant, <i>Westat, Inc.</i> |
| * M. Kovačević, <i>Statistics Canada</i> | J. Waksberg, <i>Westat, Inc.</i> |
| N. Laniel, <i>Statistics Canada</i> | C. Wu, <i>University of Waterloo</i> |
| M.D. Larsen, <i>University of Chicago</i> | Y. You, <i>Statistics Canada</i> |
| M. Latouche, <i>Statistics Canada</i> | * W. Yung, <i>Statistics Canada</i> |
| | * E. Zanutto, <i>University of Pennsylvania</i> |

Acknowledgements are also due to those who assisted during the production of the 2001 issues: H. Laplante (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Ethier, C. Larabie, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 17, No. 2, 2001

Preface	207
Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights B.K. Atrostic, Nancy Bates, Geraldine Burt, and Adriana Silberstein	209
Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century Charlotte Steeh, Nicole Kirgis, Brian Cannon, and Jeff DeWitt	227
A Theory-Guided Interviewer Training Protocol Regarding Survey Participation Robert Groves and Katherine A. McGonagle	249
Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation Elizabeth Martin, Denise Abreu, and Franklin Winters	267
The Effects of Using Administrative Registers in Economic Short Term Statistics: The Norwegian Labour Force Survey as a Case Study I. Thomsen and L.-C. Zhang	285
Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing Paul Biemer	295
Item Nonresponse in Questionnaire Research with Children Natacha Borgers and Joop Hox	321

Volume 17, No. 3, 2001

An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse Jan Pickery and Geert Loosveldt	337
The Use of Neutral Responses in Survey Questions: An Application of Multiple Correspondence Analysis Jörg Blasius and Victor Thiessen	351
Finite Sample Effects in the Estimation of Substitution Bias in the Consumer Price Index Ralph Bradley	369
Estimation of the Rates and Composition of Employment in Norwegian Municipalities Nicholas T. Longford	391
Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure Chris Moriarty and Fritz Scheuren	407
Some Statistical Problems in Merging Data Files Joseph B. Kadane	423
Book and Software Reviews	435

Volume 17, No. 4, 2001

A Neural Network Model for Predicting Time Series with Interventions and a Comparative Analysis M.D. Cubiles-de-la-Vega, R. Pino-Mejías, J.L. Moreno-Rebollo, and J. Muñoz-García	447
Understanding the Cognitive Processes of Open-Ended Categorical Questions and Their Effects on Data Quality Monica Dashen and Scott Fricker	457
What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies Robert F. Belli, Michael W. Traugott, and Matthew N. Beckmann	479
Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment Nobuaki Hoshino	499
The Delete-a-Group Jackknife Phillip S. Kott	521
Does the Model Matter for GREG Estimation? A Business Survey Example Dan Hedlin, Hannah Falvey, Ray Chambers, and Philip Kokic	527
Editorial Collaborators	545
Index to Volume 17, 2001	549

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS

TABLE DES MATIÈRES

Volume 29, No. 2, June/juin 2001

Isabel CANETTE	
Blind nonparametric regression	173
Giovanni M. MEROLA and Bovas ABRAHAM	
Dimensionality reduction approach to multivariate prediction	191
Hanfeng CHEN and Jiahua CHEN	
The likelihood ratio test for homogeneity in finite mixture models	201
M. A. TINGLEY and L. MCLEAN	
Detection of patterns in noisy time series	217
S.-Y. Claire LEI and Suojin WANG	
Diagnostic tests for bias of estimating equations in weighted regression with missing covariates	239
Anestis ANTONIADIS, Jianqing FAN and Irène GJBELS	
A wavelet method for unfolding sphere size distributions	251
Aad van der VAART and Jon A. WELLNER	
Consistency of semiparametric maximum likelihood estimators for two-phase sampling	269
Changbao WU and Randy R. SITTER	
Variance estimation for the finite population distribution function with complete auxiliary information	289
Pedro PUIG and Michael A. STEPHENS	
Goodness-of-fit tests for the hyperbolic distribution	309
Paramjit S. GILL and Tim B. SWARTZ	
Statistical analyses for round robin interaction data	321
Fatemah ALQALLAF and Paul GUSTAFSON	
On cross-validation of Bayesian models	333
Forthcoming Papers/Articles à paraître	341
Volume 29 (2001)	
Subscription rates/Frais d'abonnement	342

Volume 29, No. 3, September/septembre 2001

Peter M. HOOPER	
Flexible regression modeling with adaptive logistic basis functions	343
<i>Discussion:</i>	
Mary J. LINDSTROM: Comment 1	365
James O. RAMSAY: Comment 2	367
Nancy E. HECKMAN: Comment 3	368
Hugh A. CHIPMAN & Hong GU: Comment 4	370
<i>Rejoinder:</i>	
Peter M. HOOPER	374
Edward SUSKO, Michael J. BRONSKILL, Simon J. GRAHAM and Robert J. TIBSHIRANI	
Estimation of relaxation time distributions in magnetic resonance imaging	379
Rhonda J. ROSYCHUK and Mary E. THOMPSON	
A semi-Markov model for binary longitudinal responses subject to misclassification	395
Charmaine B. DEAN and Ying Cai MACNAB	
Modeling of rates over a hierarchical health administrative structure	405
Meehyung CHO, Nathaniel SCHENKER, Jeremy M. G. TAYLOR and Dongliang ZHUANG	
Survival analysis with long-term survivors and partially observed covariates	421
Mohan DELAMPADY, Anirban DASGUPTA, George CASELLA, Herman RUBIN and William E. STRAWDERMAN	
A new approach to default priors and robust Bayes methodology	437
John J. Spinelli	
Testing fit for the grouped exponential distribution	451
Thomas W. O'GORMAN	
Using adaptive weighted least squares to reduce the lengths of confidence intervals	459
Christophe CROUX and Catherine DEHON	
Robust linear discriminant analysis using S-estimators	473
Y. H. Steve HUANG and Longcheen HUWANG	
On the polynomial structural relationship	495
Forthcoming Papers/Articles à paraître	513
Volume 30 (2002)	
Subscription rates/Frais d'abonnement	514

Volume 29, No. 4, December/décembre 2001

Masoud ASGHARIAN and David B. WOLFSON Covariates in multipath change-point problems: modelling and consistency of the MLE	515
Nhu D. LE, Li SUN and James V. ZIDEK Spatial prediction and temporal backcasting for environmental fields having monotone data patterns	529
Louis-Paul RIVEST and Tina LÉVESQUE Improved log-linear model estimators of abundance in capture-recapture experiments	555
Edward W. FREES Omitted variables in longitudinal data models	573
Qihua WANG and J. N. K. RAO Empirical likelihood for linear regression models under imputation for missing responses	597
Shiva GAUTAM, Allan SAMPSON and Harshinder SINGH Iso-chi-squared testing of $2 \times k$ ordered tables	609
Nicolas HENGARTNER and Marten WEGKAMP Estimation and selection procedures in regression: an <i>LI</i> approach	621
Holger DETTE, Dale SONG and Weng Kee WONG Robustness properties of minimally-supported Bayesian D-optimal designs for heteroscedastic models	633
Min CHEN and Gemai CHEN A nonparametric test of conditional autoregressive heteroscedasticity for threshold autoregressive models	649
Patrice BERTAIL and Dimitris N. POLITIS Extrapolation of subsampling distribution estimators: the i.i.d. and strong mixing cases	667
Corrigenda	681
Index: Volume 29 (2001)	683
Forthcoming Papers/Articles à paraître	689
Survey methodology	690

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préférablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O, 0, 1, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

Masoud ASGHARIAN and David B. WOLFSON	515
Covariates in multipath change-point problems: modelling and consistency of the MLE	
Nhu D. LE, Li SUN and James V. ZIDEK	529
Spatial prediction and temporal backcasting for environmental fields having monotone data patterns	
Louis-Paul RIVEST and Tina LÉVESQUE	555
Improved log-linear model estimators of abundance in capture-recapture experiments	
Edward W. FREES	573
Omitted variables in longitudinal data models	
Qihua WANG and J. N. K. RAO	597
Empirical likelihood for linear regression models under imputation for missing responses	
Shiva GAUTAM, Allan SAMPSON and Harshinder SINGH	609
Iso-chi-squared testing of $2 \times k$ ordered tables	
Nicolas HENGARTNER and Marten WEGKAMP	621
Estimation and selection procedures in regression: an L_1 approach	
Holger DETTE, Dale SONG and Weng Kee WONG	633
Robustness properties of minimally-supported Bayesian D-optimal designs for heteroscedastic models	
Min CHEN and Gernai CHEN	649
A nonparametric test of conditional autoregressive heteroscedasticity for threshold autoregressive models	
Patrice BERTAIL and Dimitris N. POLYTIS	667
Extrapolation of subsampling distribution estimators: the i.i.d. and strong mixing cases	
Corrigenda	681
Index: Volume 29 (2001)	683
Forthcoming Papers/Articles à paraître	689
Survey methodology	690

Volume 29, No. 2, June/juin 2001

173	Isabel CANETTE Blind nonparametric regression
191	Giovanni M. MEROLA and Boyas ABRAHAM Dimensionality reduction approach to multivariate prediction
201	Hanfeng CHEN and Jiahua CHEN The likelihood ratio test for homogeneity in finite mixture models
217	M. A. TINGLEY and L. MCLEAN Detection of patterns in noisy time series
239	S.-Y. Claire LEI and Suojin WANG Diagnostic tests for bias of estimating equations in weighted regression with missing covariates
251	Anestis ANTONIADIS, Jianqing FAN and Irène GUBELS A wavelet method for unfolding sphere size distributions
269	Aad van der VAART and Jon A. WEILNER Consistency of semiparametric maximum likelihood estimators for two-phase sampling
289	Changbao WU and Randy R. SITTER Variance estimation for the finite population distribution function with complete auxiliary information
309	Pedro PUIG and Michael A. STEPHENS Goodness-of-fit tests for the hyperbolic distribution
321	Paramjit S. GILL and Tim B. SWARTZ Statistical analyses for round robin interaction data
333	Fatemah ALQALAF and Paul GUSTAFSON On cross-validation of Bayesian models
341	Forthcoming Papers/Articles à paraître
342	Subscription rates/Frais d'abonnement

Volume 29, No. 3, September/septembre 2001

343	Peter M. HOOPER Flexible regression modeling with adaptive basis functions
365	Mary J. LINDSTROM: Comment 1
367	James O. RAMSAY: Comment 2
368	Nancy E. HECKMAN: Comment 3
370	Hugh A. CHIPMAN & Hong GU: Comment 4
374	Peter M. HOOPER <i>Rejoinder:</i>
379	Edward SUSKO, Michael J. BRONSKIL, Simon J. GRAHAM and Robert J. TIBSHIRANI Estimation of relaxation time distributions in magnetic resonance imaging
395	Rhonda J. ROSYCHUK and Mary E. THOMPSON A semi-Markov model for binary longitudinal responses subject to misclassification
395	Charmaine B. DEAN and Ying Cai MACINAB Modelling of rates over a hierarchical health administrative structure
405	Meehyung CHO, Nathaniel SCHENKER, Jeremy M. G. TAYLOR and Dongliang ZHUANG Survival analysis with long-term survivors and partially observed covariates
421	Mohan DELAMPADY, Anirban DASGUPTA, George CASELLA, Herman RUBIN and William E. STRAWDERMAN A new approach to default priors and robust Bayes methodology
437	John J. Spinelli Testing fit for the grouped exponential distribution
451	Thomas W. O'GORMAN Using adaptive weighted least squares to reduce the lengths of confidence intervals
459	Christophe CROUX and Catherine DEHON Robust linear discriminant analysis using S-estimators
473	Y. H. Steve HUANG and Longghechen HUWANG On the polynomial structural relationship
495	Forthcoming Papers/Articles à paraître
513	Subscription rates/Frais d'abonnement
514	Volume 30 (2002)

A Neural Network Model for Predicting Time Series with Interventions and a Comparative Analysis M.D. Cubiles-de-la-Vega, R. Pino-Mejías, J.L. Moreno-Rebollo, and J. Muñoz-García	447
Understanding the Cognitive Processes of Open-Ended Categorical Questions and Their Effects on Data Quality Monica Dashen and Scott Fricker	457
What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted Nonvoters in the American National Election Studies Robert F. Belli, Michael W. Traugott, and Matthew N. Beckmann	479
Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment Nobuaki Hoshino	499
The Delete-a-Group Jackknife Phillip S. Kott	521
Does the Model Matter for GREG Estimation? A Business Survey Example Dan Hedlin, Hannah Falvey, Ray Chambers, and Philip Kokic	527
Editorial Collaborators	545
Index to Volume 17, 2001	549

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

An International Review Published by Statistics Sweden

JOURNAL OF OFFICIAL STATISTICS

Contents Volume 17, No. 2, 2001

Preface	207
Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights B. K. Atrostic, Nancy Bates, Geraldine Burt, and Adriana Silberstein	209
Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century Charlotte Steeh, Nicole Kirgis, Brian Cannon, and Jeff DeWitt	227
A Theory-Guided Interviewer Training Protocol Regarding Survey Participation Robert Groves and Katherine A. McGonagle	249
Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation Elizabeth Martin, Denise Abreu, and Franklin Winters	267
The Effects of Using Administrative Registers in Economic Short Term Statistics: The Norwegian Labour Force Survey as a Case Study I. Thomsen and L.-C. Zhang	285
Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing Paul Biemer	295
Item Nonresponse in Questionnaire Research with Children Natacha Borgers and Joop Hox	321

Volume 17, No. 3, 2001

An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse Jan Pickery and Geert Loosveldt	337
The Use of Neutral Responses in Survey Questions: An Application of Multiple Correspondence Analysis Jörg Blasius and Victor Thiessen	351
Finite Sample Effects in the Estimation of Substitution Bias in the Consumer Price Index Ralph Bradley	369
Estimation of the Rates and Composition of Employment in Norwegian Municipalities Nicholas T. Longford	391
Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure Chris Mortality and Fritz Scheuren	407
Some Statistical Problems in Merging Data Files Joseph B. Kadane	423
Book and Software Reviews	435

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 2001. Un astérisque indique que la personne a participé plus d'une fois.

- * P. Lavalée, *Statistique Canada*
 H. Lee, *Westat, Inc.*
 J. Lent, *U.S. Bureau of Transportation Statistics*
 * S. Lohr, *Arizona State University*
 W. Lu, *Simon Fraser University*
 J. Moloney, *Statistique Canada*
 G. Montanari, *University of Perugia*
 C. Perry, *NASS*
 T.E. Raghunathan, *University of Michigan*
 E. Rancourt, *Statistique Canada*
 L.-P. Rivest, *Université Laval*
 N. Schenker, *National Center for Health Statistics*
 J. Sedransk, *Case Western University*
 * A.C. Singh, *Research Triangle Institute*
 K.P. Srinath, *ABT Associates*
 B. Sutthar, *Memorial University*
 * A. Theberge, *Statistique Canada*
 R. Thomas, *Carleton University*
 S. K. Thompson, *Pennsylvania State University*
 C. Tucker, *U.S. Bureau of Labor Statistics*
 * R. Valliant, *Westat, Inc.*
 J. Waksberg, *Westat, Inc.*
 C. Wu, *University of Waterloo*
 Y. You, *Statistique Canada*
 W. Yung, *Statistique Canada*
 * E. Zanutto, *University of Pennsylvania*
- M. Axelsson, *Örebro University, Sweden*
 M. Bankier, *Statistique Canada*
 K. Brewer, *Australian National University*
 Moon Jung Cho, *U.S. Bureau of Labor Statistics*
 R. Chambers, *University of Southampton*
 * S. Chowdhury, *Westat, Inc.*
 M. P. Cohen, *U.S. Bureau of Transportation Statistics*
 G. Datta, *University of Georgia*
 P. Duchesne, *École des Hautes Études Commerciales de Montréal*
 F. Dupont, *INSEE*
 M.R. Elliott, *University of Pennsylvania*
 J.L. Eltinge, *U.S. Bureau of Labor Statistics*
 * V. M. Estevao, *Statistique Canada*
 M. Ghosh, *University of Florida*
 B. Grubard, *National Cancer Institute*
 R. Harter, *National Opinion Research Center*
 M.A. Hidiroglou, *Statistique Canada*
 B. Hultiger, *Swiss Federal Statistical Office*
 D. Jang, *Mathematica Policy Research*
 D. Kostanich, *U.S. Bureau of the Census*
 P. Kott, *NASS*
 * M. Kovacević, *Statistique Canada*
 N. Laniet, *Statistique Canada*
 M.D. Larsen, *University of Chicago*
 M. Latouche, *Statistique Canada*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 2001 : H. Laplante (Division de la diffusion) et L. Perteault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à C. Ethier, C. Larabie et D. Lemire de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, B*, 35, 61-68.

BRUNSDON, T.M. (1987). Time Series Analysis of Compositional Data. Ph.D. Thèse non publiée. University of Southampton.

BRUNSDON, T.M., et SMITH, T.M.F. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14, 3, 237-253.

DE JONG, P. (1988). The likelihood for a state space model. *Biometrika*, 75, 165-169.

DE JONG, P. (1989). Smoothing and interpolation with the state space model. *Journal of the American Statistical Society*, 84, 1085-1088.

DE JONG, P. (1991). The diffuse Kalman filter. *The Annals of Statistics*, 19, 1073-1083.

FERNANDEZ, F.J.M., et HARVEY, A.C. (1990). Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business and Economic Statistics*, 8, 1, 71-81.

GURNEY, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Social Statistics Section*, 242-257.

HARRISON, P.J., et STEVENS, C.F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, B*, 38, 205-47.

HARVEY, A.C. (1986). Analysis and generalisation of a multivariate exponential smoothing model. *Management Science*, 32, 374-380.

HARVEY, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press. Cambridge.

HARVEY, A.C. (1993). *Time Series Models*. Deuxième édition. Harvester Wheatsheaf. London.

HARVEY, A.C., et PETERS, S. (1984). Estimation Procedures for Structural Time Series Models. London School of Economics. Mimeo.

HARVEY, A.C., et SHEPHARD, N. (1993). Structural time series models. Dans *Handbook of Statistics*, (Eds. S.Maddala, C.R.Rao and H.D.Vinod). Elsevier Science Publishers, 11, 261-302.

HARVEY, A.C., et CHUNG, C. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, A*, 163, Part 3, 303-339.

IBGE (1980). Metodologia da Pesquisa Mensal de Emprego 1980. Relatórios Metodológicos. Fundação Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro.

JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, B*, 42, 221-226.

KOOPMAN, S.J., HARVEY, A.C., DOORNIK, J.A., et SHEPHARD, N. (1995). *STAMP 5.0 - Structural Time Series Analyser, Modeler and Predictor*. Chapman & Hall. London.

MITNIK, S. (1991). Deviation of the unconditional state covariance matrix for exact-likelihood estimation of ARMA models. *Journal of Economic Dynamics and Control*, 15, 731-740.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-177.

PFEFFERMANN, D., BURCK, L., et BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Proceedings of the International Symposium on Analysis of Data in Time*, 43-55.

PFEFFERMANN, D., et BURCK, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales. *Techniques d'enquête*, 16, 229-249.

PFEFFERMANN, D., et BLEUER, S.R. (1993). Modélisation conjointe robuste de séries de données sur l'activité pour de petites régions. *Techniques d'enquête*, 19, 159-174.

PFEFFERMANN, D., BELL, P. et SIGNORELLI, D. (1996). Labour force trend estimation in small areas. *Proceedings of the Annual Research Conference, Bureau of the Census*, 407-431.

PFEFFERMANN, D., FEBER, M. et SIGNORELLI, D. (1998). Estimation of autocorrelations of survey errors with applications to trend estimation in small samples. *Journal of Business and Economics Statistics*, 16, 339-348.

QUINTANA, J.M., et WEST, M. (1988). Time series analysis of compositional data. *Journal of Bayesian Statistics*, (Eds. J.H. Bernardo, M.A. Degroot et A.F.M. Smith). Oxford University Press, 3.

REINSEL, G.C. (1993). *Elements of Multivariate Time Series Analysis*. Springer-Verlag.

SAS INSTITUTE INC. (1995). *SAS/IML Software : Changes and Enhancements through Release 6.11*. SAS institute Inc. Cary, NC.

SCOTT, A.J., et SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

SCOTT, A.J., SMITH, T.M.F. et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.

SHEPHARD, N.G., et HARVEY, A.C. (1989). Tracking the Level of Support for the Parties During General Election Campaigns. Mimeo. Dept. of Statistics, London School of Economics.

SILVA, D.B.N. (1996). Modelling Compositional Time Series From Repeated Surveys. Thèse non publiée Ph.D. University of Southampton, UK.

SINGH, A.C., et ROBERTS, G.R. (1992). State space modelling of cross-classified time series of counts. *International Statistical Review*, 60, 321-335.

SMITH, T.M.F., et BRUNSDON, T.M. (1986). Time Series Methods for Small Areas. Rapport non publié. University of Southampton.

TIAO, G.C., et BOX, G.E.P. (1981). Modelling multiple time series with applications. *Journal of the American Statistical Association*, 76, 802-816.

TILLER, R.B. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 16-25.

TILLER, R.B. (1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 2, 149-166.

WALLIS, F. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, 8, 115-23.

WEI, W.W.S. (1993). *Time Series Analysis - univariate and multivariate methods*. Addison-Wesley.

WEST, M., et HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

Nous avons obtenu les estimations de la tendance axée sur le modèle par ajustement du modèle structural de base défini pour le processus d'émission du signal en modélisant la variation de l'erreur d'échantillonnage sous forme de modèle VARMA(1,1). Nous avons comparé ces estimations à celles produites par la méthode X-11 bien connue. La figure 3 montre les tendances produites pour la série de taux de chômage par les deux méthodes ainsi que les estimations obtenues par ajustement d'un modèle structural de base du type qui ne tient pas compte de la variation de l'erreur de l'échantillonnage.

La tendance produite par notre modèle est plus lisse, ce qui donne à penser que la méthode axée sur le modèle permet d'éliminer les fluctuations causées par les erreurs d'échantillonnage. En outre, nous avons également obtenu les estimations axées sur le modèle pour les effets saisonniers des compositions originales par la méthode de décomposition multivariée qui tient compte de deux caractéristiques importantes des données, à savoir les contraintes compositionnelles et l'existence d'erreurs d'échantillonnage.

Nous proposons une méthode espace-état pour modéliser les séries chronologiques compositionnelles produites d'après les données d'enquêtes répétées. La caractéristique importante de la méthode proposée tient au fait qu'elle fournit des prédictions et des estimations bornées du signal des paramètres d'une composition qui satisfont la contrainte de somme unitaire, tout en tenant compte de l'effet des erreurs d'échantillonnage. Pour y arriver, nous procédons au mappage des compositions de l'espace simplexe sur l'espace réel grâce à la transformation additive par le logarithme du ratio, nous modélisons les données transformées au moyen de modèles espace-état multivariés, puis nous appliquons la transformation additive logarithmique pour obtenir les estimations à l'échelle originale.

Nos travaux empiriques portant sur les données de l'Enquête sur la population active du Brésil démontrent l'utilité de cette méthode de modélisation dans une situation d'enquête réelle, prouvant qu'il est possible de modéliser le système multivarié et d'obtenir des estimations pour toutes les composantes pertinentes. Ces travaux empiriques montrent aussi que l'utilisation d'un modèle qui tient compte explicitement des erreurs d'échantillonnage produit une tendance et des saisonniers fixes plus lisses que les estimations produites par le modèle X-11. En outre, comme les estimateurs axés sur le modèle combinent les données d'enquêtes passées et courantes, les écarts-types des estimations obtenues sont en général plus faibles que les écarts-types des estimateurs fondés sur le plan de sondage, comme l'a montré Silva (1996, chapitre 8).

L'un des inconvénients de la méthode proposée tient au fait que, même si les régions de confiance du vecteur compositionnel original peuvent être construites d'après les estimations axées sur le modèle par application de la loi de

6. CONCLUSIONS

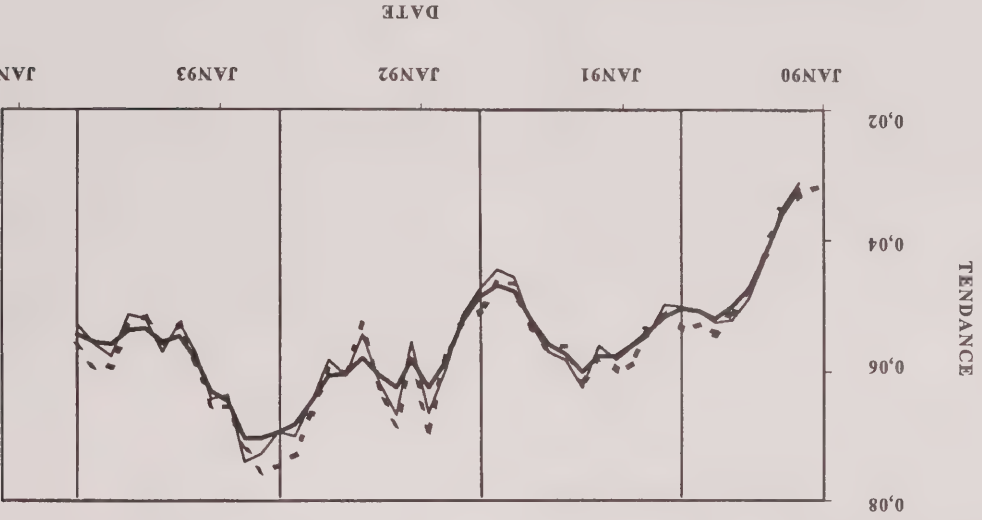
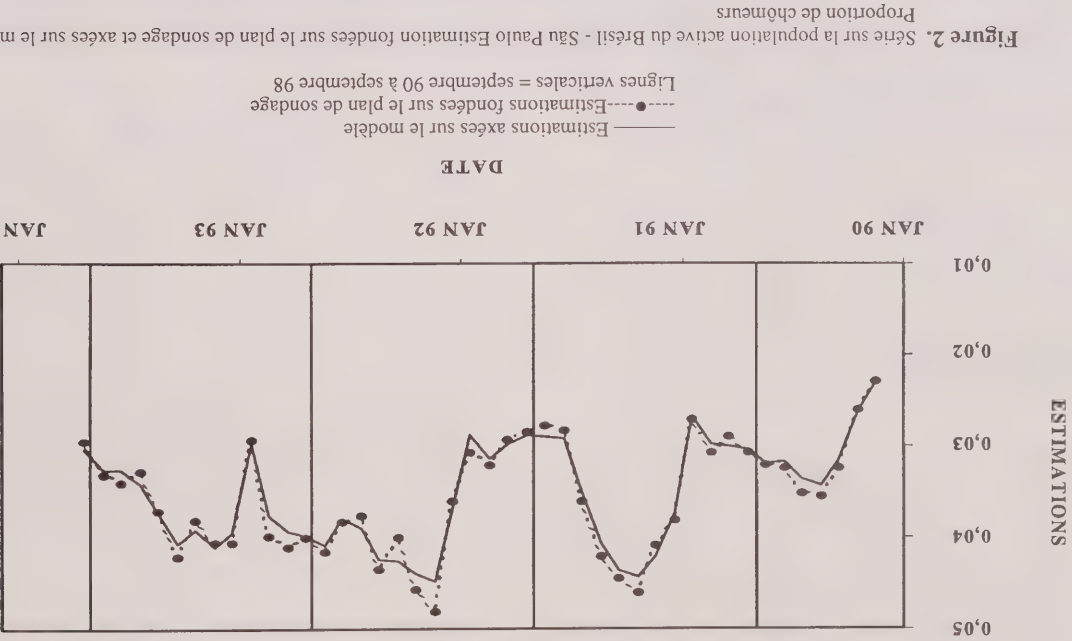
L'un des grands avantages de la méthode de modélisation proposée tient au fait que, dans le cas de la formation d'un espace-état, on dispose d'une vaste gamme de modèles pour représenter les processus multivariés de signal et de bruit. L'application de la méthode à divers ensembles de données serait souhaitable. D'autres travaux empiriques devraient aussi inclure des situations où les compositions se trouvent dans un espace simplexe ayant des dimensions supérieures à deux et(ou) s'approchent des bornes de l'intervalle [0,1]. En outre, l'application de la méthode à des données simulées, pour lesquelles on connaît les « vrais » modèles sous-jacents, donnerait une meilleure idée de la performance de la méthode de modélisation. Les modèles considérés ici peuvent aussi être généralisés afin d'y intégrer les effets du biais dû au groupe de renouvellement et certaines variables explicatives.

REMERCIEMENTS

Les présents travaux ont été financés par CAPES-Brésil et IBGE-Brésil et par une bourse de l'Economic and Social Research Council du Royaume-Uni octroyée dans le cadre de son Analysis of Large and Complex Datasets Programme. Les auteurs remercient les examinateurs et le professeur Danny Pfeffermann de leurs suggestions qui leur ont permis d'apporter de nombreuses améliorations à l'article. Ils remercient aussi M. Harold Mantel de son soutien lors de la préparation de la version finale.

BIBLIOGRAPHIE

- AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. New York : Chapman and Hall.
- AITCHISON, J., et SHEN, S.M. (1980). Logistic-Normal distributions : some properties and uses. *Biometrika*, 67, 261-272.
- BELL, W.R., et HILLMER, S.C. (1990). Estimation dans les enquêtes à passages répétées au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.
- BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Dans Handbook of Statistics*, (Eds. P.R. Krishnaiah et C.R. Rao). Elsevier Science, 6, 187-211.
- BINDER, D.A., et DICK, J. P. (1989). Enquêtes répétées - modélisation et estimation. *Techniques d'enquête*, 15, 31-48.
- BINDER, D.A., BLEUER, S.R. et DICK, J.P. (1993). Time series methods applied to survey data. *Proceedings of the 49^e International Statistical Institut Session*, 1, 327-344.



du signal d'origine $\theta_{m,t|T}$, telles que $L_{m,t|T}$ et $S_{m,t|T}$. Cependant, si l'on suppose pour $\{\theta_{m,t|T}\}$, un modèle multiplicatif sans composantes irrégulières, de sorte que :

$$\theta_{1t} = L_{1t} S_{1t}, \theta_{2t} = L_{2t} S_{2t}, \theta_{3t} = L_{3t} S_{3t}, \quad (19)$$

où $L_{m,t}$ et $S_{m,t}$, pour $m = 1, 2, 3$ représentent les composantes de tendance et de saisonnalité des signaux inobservables, alors, si nous appliquons une transformation additive par logarithme du ratio à $\theta_{m,t}$, nous obtenons :

$$\log(\theta_{m,t}/\theta_{3,t}) = \log\left(\frac{L_{m,t} S_{m,t}}{L_{3,t} S_{3,t}}\right) = \log\left(\frac{L_{m,t}}{L_{3,t}}\right) + \log\left(\frac{S_{m,t}}{S_{3,t}}\right), \quad m = 1, 2. \quad (20)$$

Cette expression peut être réécrite sous la forme :

$$\theta_{m,t}^* = L_{m,t}^* + S_{m,t}^*, \quad (21)$$

avec $L_{m,t}^* = \log(L_{m,t}/L_{3,t})$ et $S_{m,t}^* = \log(S_{m,t}/S_{3,t})$. Donc, l'utilisation d'un modèle structurel de base pour $\{\theta_{m,t}^*\}$ correspond au cas où le modèle sous-jacent de $\{\theta_{m,t}\}$ pose le signal original en ses composantes de tendance et de saisonnalité de façon multiplicative. Pour calculer les estimations, filtrées ou lissées, de $L_{m,t}^*$, notons que :

$$\exp(L_{1t}^*) = L_{1t}/L_{3t}, \exp(L_{2t}^*) = L_{2t}/L_{3t}. \quad (22)$$

Pour récupérer $L_{1t}^*, L_{2t}^*, L_{3t}^*$, dans (22), il est nécessaire de supposer qu'il existe une relation explicite entre ces composantes inobservables fondées sur le modèle (19). De la sorte, nous pouvons ajouter une troisième équation au système dans (22) et obtenir une estimation des composantes de la série originale. Notons qu'il s'agit d'un système de deux équations à trois inconnues. Dans ce cas, il est assez naturel de supposer que la somme des composantes de niveau sur les séries est égale à un, étant également comprise entre les bornes zéro et un. Par conséquent, pour la série originale, nous pouvons obtenir les estimations de la tendance en résolvant :

$$\left\{ \begin{array}{l} \exp(L_{1t}^*) = L_{1t}/L_{3t}, \\ \exp(L_{2t}^*) = L_{2t}/L_{3t}, \\ L_{1t} + L_{2t} + L_{3t} = 1, \end{array} \right. \quad (23a)$$

qui produit

ment aigus de la série aient été lissés.

La figure 2 présente les estimations fondées sur le plan de sondage et les estimations axées sur le modèle de la portion de chômeurs, pour la période allant de janvier 1989 à septembre 1993. Les estimations indépendantes du modèle sont les estimations lissées fondées sur l'ensemble de données pour toute la période couverte par l'échantillon. Comme le montre le graphique, les estimations du signal se comportent de façon comparable aux estimations fondées sur le plan de sondage, quoique certains points de renverse-

ment aigus de la série aient été lissés. La figure 2 présente les estimations fondées sur le plan de sondage et les estimations axées sur le modèle de la portion de chômeurs, pour la période allant de janvier 1989 à septembre 1993. Les estimations indépendantes du modèle sont les estimations lissées fondées sur l'ensemble de données pour toute la période couverte par l'échantillon. Comme le montre le graphique, les estimations du signal se comportent de façon comparable aux estimations fondées sur le plan de sondage, quoique certains points de renverse-

$$R_t' = \frac{\theta_{1t} + \theta_{2t}}{\theta_{1t}} = \frac{1 + \frac{\theta_{1t}}{\theta_{2t}}}{1} = \left(\frac{\theta_{1t}}{\theta_{2t}} + 1 \right)^{-1}. \quad (24)$$

D'après le modèle (11), nous pouvons obtenir les estimations de la tendance pour le taux de chômage en remplaçant simplement $\theta_{m,t}$ par $L_{m,t}$, $m = 1, 2$, dans l'équation 24. En somme, la méthodologie développée à la présente section produit des estimations du signal (et de la tendance) qui sont bornées entre zéro et un et satisfont la contrainte de somme unitaire. Elle fournit aussi des estimations des composantes saisonnière et tendance de la série comprenant les ratios des proportions originales, c'est-à-dire une caractéristique fort utile.

Dans les cas des enquêtes sur la population active, une question importante est celle de l'estimation des séries de taux de chômage (par opposition aux proportions de chômeurs) ainsi que la production des chiffres de saison-

nalises correspondants. Rappelons que θ_{1t} et θ_{2t} représentent les proportions inconnues de chômeurs et de personnes occupées dans la population, respectivement. Partant de ces proportions, nous définissons le taux de chômage inconnu à la période t comme étant

$$S_{m,t|T}^* = \theta_{m,t|T}/L_{m,t|T}^*, \quad m = 1, 2, 3.$$

Comme il n'existe aucune composante irrégulière dans les estimations de la tendance dans (23). Par conséquent, nous obtenons les estimations lissées de la tendance de la série originale de proportions en appliquant la transformation logarithmique additive à $L_{m,t|T}^*$. Conséquemment, nous pouvons calculer les estimations des composantes saisonnières des proportions originales sous la forme :

$$L_{m,t}^* = \frac{\exp(L_{m,t}^*)}{\sum_{k=1}^K \exp(L_{k,t}^*)}, \quad m = 1, 2; \quad L_{3,t}^* = \frac{1 + \sum_{k=1}^K \exp(L_{k,t}^*)}{1}. \quad (23b)$$

représenter le processus transformé d'erreurs d'échantillonnage. En l'occurrence, nous avons choisi le modèle VARMA(1,1) parce qu'il produit des erreurs-types plus faibles pour les estimations du taux de chômage. Nous avons estimé les paramètres de ce modèle d'après la relation entre la fonction de covariances croisées et les matrices paramétriques présentées dans Wei (1993, pages 346 et 347). Le modèle VARMA(1,1) ajusté pour $\{e_t^*\}$ est donné par :

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} 0,7347 & 0,2414 \\ -0,9224 & -0,2072 \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} - \begin{bmatrix} 0,3162 & 0,2590 \\ -0,7666 & -0,2749 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}, \quad (18)$$

avec

$$\sum_a = \begin{bmatrix} 0,0001723 & 0,0003476 \\ 0,0003476 & 0,0051660 \end{bmatrix}.$$

Après avoir produit le modèle combiné pour les estimations d'enquêtes transformées sous forme d'espace-état, nous pouvons utiliser le filtre de Kalman pour obtenir des estimations filtrées et lissées des composantes inobservables. Notons que, pour les erreurs d'échantillonnage transformées (équation 18), l'estimation du modèle a été réalisée sans le filtre de Kalman. L'application de ce dernier nécessite l'estimation des hyperparamètres inconnus (les covariances) du vecteur d'état initial et de la matrice des covariances respectives. Si nous supposons que les perturbations $\eta_{(t)}^*$, a_t^* et I_t^* obéissent à la loi de distribution normale, la fonction du logarithme du rapport des vraisemblances des observations (transformées) peut être exprimée par décomposition de l'erreur de prédiction (pour plus de précision, voir Harvey 1989). Nous avons estimé les covariances du modèle par la méthode du maximum de vraisemblance, en appliquant une technique de quasi-optimisation de Newton. Nous avons mis au point un programme informatique afin de mettre en œuvre la méthode de maximisation fondée sur la routine d'optimisation NLPQN de SAS-IML.

Pour initialiser le filtre de Kalman, nous avons utilisé une combinaison de conditions a priori diffuses et propres. Suivant cette méthode, nous avons initialisé les composantes non stationnaires ($\alpha^{(0)}$) du vecteur d'états avec des variances d'erreurs très grandes et nous avons donné aux composantes respectives du vecteur d'états initial une valeur nulle. Les composantes stationnaires (e_{1t}^*, e_{2t}^*) ont été initialisées d'après la moyenne et la variance inconditionnelles correspondantes.

Lors de l'ajustement du modèle, les matrices des covariances estimatives obtenues pour la pente et les composantes saisonnières étaient très faibles et leur valeur a pu

Table 1
Estimations des hyperparamètres et des erreurs-types

Modèle	$\hat{\Sigma}_t \times 10^{-4}$ (2)	$\hat{\Sigma}_t = \hat{\Sigma}_s = \hat{\Sigma}_l$
BSM + VARMA (1,1)	$\begin{bmatrix} 2,78 & 0,12 \\ 0,91 & 1,95 \\ 87,0 & (3,55) \end{bmatrix}$	(1)
(1) Modèle local de niveau avec décalage et événements saisonniers fixes pour le signal.	(2) L'élément triangulaire supérieur contient une corrélation.	

Pour évaluer la performance du modèle, nous avons comparé les distributions empiriques des résidus normalisés à la distribution normale type pour vérifier l'hypothèse selon laquelle les innovations $(v_t - v_{t-1}^{(1)})$ sont des écarts aléatoires normaux. L'examen des représentations graphiques normales correspondantes ne révèle aucun écart par rapport à la normalité. En outre, nous avons calculé les autocorrélations des innovations, dont la valeur était proche de zéro, afin de valider encore davantage le modèle.

Nous calculons les prédictions pour y_{mt}^* et les estimations de $\theta_{t|T}^*$ par application de la transformation additive logarithmique (équation 5) aux prédictions $v_{t|T-1}^*$ et aux estimations lissées $\theta_{t|T}^*$ pour la série et le signal transformés, respectivement. Cette transformation permet le mappage de ces estimations dans S_2^* , assurant qu'elles satisfassent les contraintes d'absence de bornes.

Malheureusement, alors que l'on peut obtenir $L_{t|T}^*$ et $S_{t|T}^*$ d'après $\theta_{t|T}^*$, il n'est pas simple d'obtenir les estimations pour les composantes structurelles inobservables

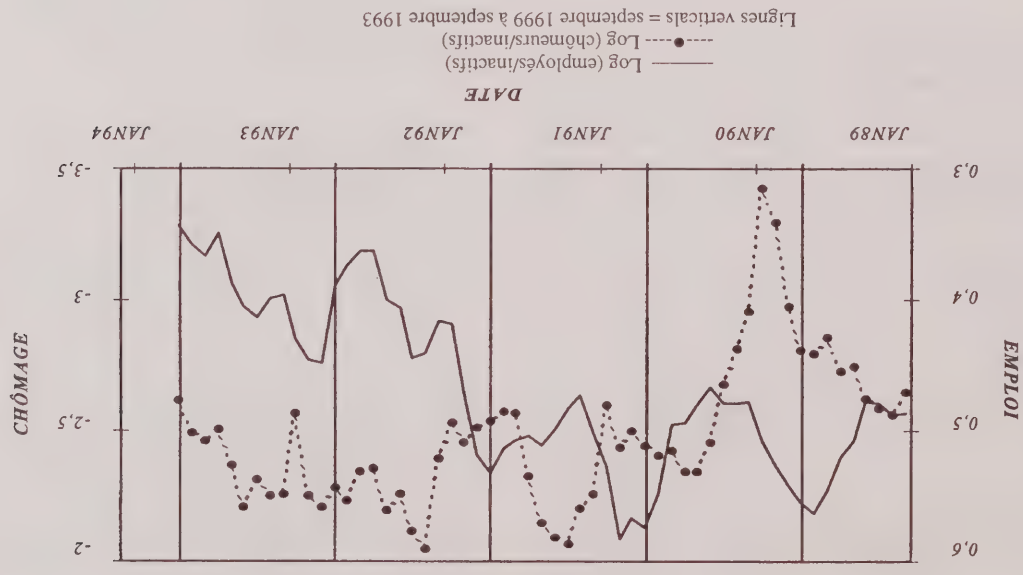


Figure 1. Série sur la population active du Brésil - São Paulo Compositions transformées

au moyen de données de chaque cycle particulier de l'enquête et d'estimateurs types. Pour chaque mois, nous avons obtenu deux ensembles d'estimations. Les estimations directes d'échantillon, calculées d'après les données complètes recueillies lors d'un mois particulier et quatre estimations élémentaires, chacune fondée sur des données provenant d'un seul groupe de renouvellement. Les estimations de panel permettent d'évaluer les autocorrélations des erreurs d'échantillonnage et de préciser plus facilement le modèle de série chronologique pour ces erreurs.

Dans la présente étude, la composition observée correspond $M+1 = 3$ composantes et la série chronologique correspond à la série de vecteurs $y_t = (y_{1t}, y_{2t}, y_{3t})'$, où :

- y_{1t} est la proportion estimative de personnes au chômage durant le mois t ;
- y_{2t} est la proportion estimative de personnes employées durant le mois t ;
- y_{3t} est la proportion estimative de personnes qui ne font pas partie de la population active au mois t .

Le modèle de l'EPAB doit intégrer les caractéristiques spéciales des données. En premier lieu, il s'agit d'une série chronologique compositionnelle appartenant au simplexe S^2 à chaque période t . En deuxième lieu, les séries chronologiques sont sujettes à des erreurs d'échantillonnage. Suivant la méthode exposée à la section 2, nous commençons par mapper la composition sur \mathbb{R}^2 au moyen de la transformation additive par logarithme du ratio en prenant y_{3t} comme catégorie de référence. Comme y_t est un vecteur d'estimations d'échantillon, il peut être modélisé comme dans l'équation 1 et le modèle vectoriel de la série transformée est donné par l'équation 9. Puis, la composition transformée est modélisée au moyen d'un modèle multivarié espace-état qui tient compte des autocorrélations entre les erreurs d'échantillonnage. Enfin, les estimations fondées sur le modèle sont transformées pour retourner à l'espace-état de départ. La figure 1 montre la série de compositions transformées.

Le modèle appliqué aux estimations d'échantillon transformées y_t est composé d'un modèle bivarie du signal transformé θ_t , décrivant comment les quantités de population transformées évoluent en fonction du temps et d'un modèle bivarie représentant la relation chronologique entre les erreurs d'échantillonnage transformées e_t . Nous supposons que le processus transformé d'émission du signal $\{\theta_t\}$ suit un modèle structurel de base bivarie (équation 11) tel que décrit à la section 3. Comme nous l'avons mentionné plus haut, nous utilisons un modèle VARMA pour représenter la série d'erreurs d'échantillon. La structure de corrélation des erreurs d'échantillonnage transformées a été estimée d'après les pseudo-erreurs transformées, comme dans le cas de l'équation 16. En outre, les estimations des matrices des corrélations partielles des retards pour $\{e_t\}$ sont calculées au moyen d'un algorithme récursif fourni par Wei (1993, pages 359 à 362). Nous avons développé un programme SAS-ML qui donne les représentations schématiques correspondantes (Tiao et Box 1981) ainsi qu'un test statistique pour établir plus facilement l'ordre du processus vectoriel. La forme des matrices des corrélations et les résultats du test statistique disponibles dans Silva (1996), indiquent que l'on pourrait utiliser un modèle VAR(1), VAR(2) ou VARMA(1,1) pour

5. MODÉLISATION DE SÉRIES CHRONOLOGIQUES COMPOSITIONNELLES DANS LE CAS DE L'ENQUÊTE SUR LA POPULATION ACTIVE DU BRÉSIL

L'Enquête sur la population active du Brésil (EPAB) est conçue en vue de recueillir des données mensuelles sur l'emploi, le nombre d'heures de travail, le niveau de scolarité et la rémunération, ainsi que d'autres renseignements géographiques. Elle permet de classer les répondants, âgés de 15 ans et plus, en fonction de leur situation d'activité durant la semaine qui a précédé l'entrevue en trois grands groupes, à savoir les personnes occupées, les chômeurs et les personnes qui ne font pas partie de la population active, conformément aux définitions de l'Organisation internationale du travail (OIT). L'enquête a pour cible la population des six grandes régions métropolitaines du pays. L'EPAB est une enquête par sondage à deux degrés pour laquelle les unités primaires d'échantillonnage (UPF) sont les secteurs de dénombrement (SD) et les unités de deuxième degré sont les ménages. Les unités primaires d'échantillonnage sont sélectionnées avec probabilité proportionnelle à la taille, puis un nombre fixe de ménages sont sélectionnés par échantillonnage systématique dans chaque SD échantillonné. Tous les membres des ménages sélectionnés sont recensés. Les unités primaires d'échantillonnage restent les mêmes pour une période d'environ 10 ans (comme dans un échantillon principal). De nouvelles unités primaires d'échantillonnage sont sélectionnées lorsque des renseignements provenant d'une nouvelle population de recensement deviennent disponibles.

En outre, L'EPAB est une enquête avec renouvellement de panels. Quel que soit le mois, l'échantillon est composé de quatre groupes de renouvellement contenant des ensembles s'excluant mutuellement d'unités primaires d'échantillonnage. Le processus de renouvellement s'applique aux panels d'unités secondaires d'échantillonnage (ménages). À l'intérieur de chaque groupe de renouvellement, un panel de ménages fait partie de l'échantillon pendant quatre mois successifs, en est retiré pendant les huit mois suivants, puis est de nouveau échantillonné pour une autre période de quatre mois successifs. Chaque mois, un panel est retiré de l'échantillon. Le panel qui le remplace peut être nouveau ou avoir déjà été observé pendant la période des quatre premiers mois. Notons que le plan de rotation 4-8-4 cause une structure de corrélation complexe des erreurs d'échantillonnage au cours du temps et qu'il existe un chevauchement de 75 % entre les échantillons de deux mois successifs. Les travaux empiriques ont été réalisés au moyen de données provenant de la région métropolitaine de São Paulo pour la période allant de janvier 1989 à septembre 1993 (57 observations). Les quantités étudiées sont les proportions de personnes occupées, de chômeurs et de personnes qui ne font pas partie de la population active, ainsi que le taux de chômage. En nous servant des observations mensuelles individuelles, nous avons calculé la série d'estimations d'échantillon et les erreurs-types respectives

$$\hat{e}_{i \cdot (k)}^t = (\hat{e}_{1 \cdot (k)}^t, \dots, \hat{e}_{M^t \cdot (k)}^t)' = \mathbf{v}_{i \cdot (k)}^t - \mathbf{v}_{i \cdot (k)}^t = (\mathbf{v}_{1 \cdot (k)}^t, \dots, \mathbf{v}_{M^t \cdot (k)}^t - \mathbf{v}_{M^t \cdot (k)}^t)', \quad (16)$$

où $\mathbf{v}_{i \cdot (k)}^t = 1/K \sum_{k=1}^K \mathbf{v}_{i \cdot (k)}^t$. Notons, en outre, que $\hat{e}_{i \cdot (k)}^t = \mathbf{e}_{i \cdot (k)}^t - \mathbf{e}_{i \cdot (k)}^t$. Les équations (14) et (15) montrent clairement que le cadre proposé par Pfeffermann, Bell et Signorelli (1996) s'applique aussi au modèle transformé. Nous pouvons obtenir les matrices d'autocorrélations croisées des erreurs d'échantillonnage transformées en calculant la moyenne des matrices des covariances croisées des pseudo-erreurs transformées comme suit (pour plus de précision, consulter Silva 1996, chapitre 7) :

$$\mathbf{P}_{e \cdot (h)}^t = \left[\sum_{k=1}^K \mathbf{D}_{e \cdot (k)}^t \right]^{-1/2} \left[\sum_{k=1}^K \mathbf{F}_{e \cdot (k)}^t(h) \right] \left[\sum_{k=1}^K \mathbf{D}_{e \cdot (k)}^t \right]^{-1/2}, \quad (17)$$
$$\mathbf{F}_{e \cdot (h)}^t(h) = \text{COV}(\mathbf{e}_{i \cdot (k)}^{t-h}, \mathbf{e}_{i \cdot (k)}^t) = E(\mathbf{e}_{i \cdot (k)}^{t-h} \mathbf{e}_{i \cdot (k)}^t),$$
$$\{\mathbf{F}_{e \cdot (h)}^t(h)\}_{m_f} = \text{COV}(\mathbf{e}_{m \cdot (k)}^{t-h}, \mathbf{e}_{m_f \cdot (k)}^t) = \gamma_{m \cdot (k)}^{e_{m_f}}(h)$$

Après avoir estimé les matrices des corrélations $\mathbf{P}_{e \cdot (h)}^t, h = 1, 2, \dots$ nous pouvons sélectionner un modèle VARMA pour représenter le processus transformé d'erreurs d'enquêtes et calculer des estimations des matrices paramétriques respectives, à condition que soient disponibles les séries de pseudo-erreurs transformées. Puis, comme nous l'avons décrit à la section 3, nous pouvons définir un modèle espace-état pour représenter le signal et les erreurs d'échantillonnage transformés, et nous pouvons utiliser les équations du filtre de Kalman pour produire des estimations filtrées et lissées pour les composantes inobservables. L'application du filtre de Kalman nécessite l'estimation des hyperparamètres inconnus (les matrices des covariances des covariances respectives $\sum_1^t, \sum_r^t, \sum_s^t, \sum_a^t$ et celle du vecteur d'état initial et ses matrices des covariances respectives. Maintenant que nous abordons la question de savoir comment modéliser les estimations d'enquête dans un cadre compositionnel et comment préciser le modèle de série chronologique à appliquer aux erreurs d'échantillonnage transformées, nous présentons à la section suivante les résultats d'une étude empirique fondée sur des données compositionnelles provenant de l'Enquête sur la population active du Brésil.

section 4, nous considérons l'estimation du modèle et à la section 5, nous illustrons son application au moyen de données de l'Enquête sur la population active du Brésil.

3. MODÉLISATION DE LA SÉRIE TRANSFORMÉE

Notre méthode se fonde sur l'hypothèse que la série transformée $v^m = a^m(y^m)$ a la structure signal plus bruit de l'équation 9. Nous proposons des modèles structurels de série chronologique pour $\{\theta^m_t\}$, comme ceux décrits par Harvey (1989), et des modèles vectoriels ARMA (Tiao et Box 1981) pour $\{e^m_t\}$.

Nous supposons que le processus transformé d'émission du signal $\{\theta_i^m\}$ obéit au modèle structurel multivariée de base et que chacune des composantes $\{\theta_i^{m_l}\}$ obéit à un modèle structurel de série chronologique de base (MSB) pour lequel les paramètres peuvent éventuellement différer selon la série. La structure de corrélation des perturbations du système rend compte des relations transversales entre les séries. Le modèle utilisé pour $\{\theta_i^{m_l}\}$, $m = 1, 2, \dots, M$ est alors donné par :

$$\left. \begin{aligned} {}^e{}^{\mathcal{W}}\mathbf{u}_{(s)} + \Gamma^{-1'}{}^{\mathcal{W}}S \sum_{11}^{1=f} &= {}^{\mathcal{W}}S_* \\ {}^e{}^{\mathcal{W}}\mathbf{u}_{(r)} + \Gamma^{-1'}{}^{\mathcal{W}}R &= {}^{\mathcal{W}}R_* \\ {}^e{}^{\mathcal{W}}\mathbf{u}_{(l)} + \Gamma^{-1'}{}^{\mathcal{W}}R_* + \Gamma^{-1'}{}^{\mathcal{W}}I &= {}^{\mathcal{W}}I_* \\ {}^e{}^{\mathcal{W}}I_* + {}^{\mathcal{W}}S_* + {}^{\mathcal{W}}I &= {}^{\mathcal{W}}\theta \end{aligned} \right\} \quad (OI)$$

où $L_{m_i}^{m_i}$ représente la composante de tendance/niveau du signal $\theta_{m_i}^{m_i}$, $R_{m_i}^{m_i}$ représente la variation correspondante du niveau, $S_{m_i}^{m_i}$ représente la composante saisonnière et $I_{m_i}^{m_i}$ représente une composante irrégulière. Pour chaque composante, nous supposons que les perturbations $\eta_{m_i}^{(s)}$, $\eta_{m_i}^{(i)}$, $\eta_{m_i}^{(s)}$ et l'irrégulier $I_{m_i}^{m_i}$, sont des écarts aléatoires normaux et mutuellement non corrélés dont la moyenne est nulle et dont les variances sont $\sigma_{m_i}^2$, $\sigma_{m_i}^2$, $\sigma_{m_i}^2$, respectivement. Autrement dit, les $M \times 1$ perturbations vectorielles $\eta_{(1)}^{(1)}$, $\eta_{(1)}^{(s)}$ et $I_{(1)}^{m_i}$ sont mutuellement non corrélées pour toutes les périodes de référence. En outre, nous supposons que les irréguliers $I_{m_i}^{m_i}$, $I_{(i-h)}^{(i-h)}$ avec $m \neq j$, $h = \dots, -2, -1, 0, 1, 2, \dots$ sont corrélés lorsque $h = 0$, mais ne le sont pas pour $h \neq 0$ et que $I_{(1)}^{m_i}$ a une matrice des covariances $\sum_{(a)} I_{(a)}^{m_i}$. Il en est de même pour les perturbations du système $\eta_{(a)}^{(a)}$, $\eta_{(j-(i-h))}^{(i-h)}$, $a = l, r, s$, qui sont également corrélés lorsque $h = 0$, mais ne le sont pas pour $h \neq 0$, avec les matrices des covariances $\sum_{(s)} I_{(s)}^{m_i}$, $\sum_{(r)} I_{(r)}^{m_i}$, $\sum_{(l)} I_{(l)}^{m_i}$. À chaque période i , la structure de corrélation des composantes de la composition se réduit à une matrice diagonale par blocs

$$(II) \quad \left. \begin{aligned} {}^i u_{(0)} \mathcal{D} + {}^{I-i} p_{(0)} \mathcal{L} &= {}^i p \\ {}^i I + {}^i p_{(0)} H &= {}^i \theta \end{aligned} \right\}$$

où les blocs sont $\sum_{i=1}^p \sum_{j=1}^p \sum_{s=1}^S$. Notons que la relation entre les scores se fait par la voie des éléments non diagonaux non nuls des matrices des covariances des perturbations. Le modèle multivarié (10) donne pour $\{\theta'_i\}$ correspond à la formulation espace-état suivante :

$${}^{\epsilon}W_I \otimes \begin{bmatrix} \varepsilon \times 10^7 \mathbf{0} \\ \dots \dots \\ {}^{\varepsilon}I \end{bmatrix} = {}_{(b)}\mathcal{G}$$

$${}^{\mathcal{W}}I \otimes \left[\begin{array}{ccccccc} 0 & 1 & \cdots & 0 & 0 & : & \\ \vdots & \vdots & & \vdots & \vdots & : & \\ 0 & 0 & \cdots & 1 & 0 & : & \mathbf{0}_{11 \times 2} \\ 0 & 0 & \cdots & 0 & 1 & : & \\ 1 & 1 & \cdots & 1 & 1 & : & \\ \vdots & \vdots & \cdots & \vdots & \vdots & : & \vdots \end{array} \right] = {}_{(\Theta)}\mathcal{I}$$

Nous supposons que le processus transformé d'erreurs $\{e_i'\}$ correspond à un processus vectoriel autorégressif à moyennes mobiles (VARMA) M -dimensionnel défini par $\Phi(B)e_i' = \Theta(B)a_i$, avec le vecteur de moyenne $E(e_i') = 0$ et

$${}^d_B\Phi - \dots - {}^1_B\Phi - I = (B)\Phi$$

où les matrices des coefficients $\Phi^1, \dots, \Phi^d, \Theta^1, \dots, \Theta^b$ et \mathbf{a}' est un vecteur aléatoire de bruit blanc M -dimensionnel dont la moyenne est nulle et la structure de covariance est :

$$\left. \begin{array}{l} 0 \neq y \quad \mathbf{0} \\ 0 = y \quad v \Sigma \end{array} \right\} = (y^{-1} v' v) E$$

des erreurs d'échantillonnage, $y_{m,t}$ peut être décomposée

comme suit :

$$y_{m,t} = \theta_{m,t} + e_{m,t}, \quad m = 1, \dots, M + 1, \quad (2)$$

où $\theta_{m,t}$ est la proportion de population inconnue qui, en

principe, obéit à un modèle de série chronologique et $e_{m,t}$

est l'erreur d'échantillonnage. Si l'on considère les $M+1$

séries simultanément, l'équation (2) peut s'écrire sous

forme vectorielle comme dans l'équation 1. En outre, nous

supposons que

$$\sum_{m=1}^M \theta_{m,t} = \sum_{m=1}^M y_{m,t} = 1 \quad \forall t, \quad (3)$$

ce qui sous-entend que $\sum_{m=1}^M e_{m,t} = 0, \quad \forall t$.

Une série chronologique compositionnelle est une série

de vecteurs $y_t = (y_{1,t}, \dots, y_{M+1,t})'$ qui appartiennent chacun

à S^M . Aitchison (1986) a décrit les difficultés que pose

l'application des méthodes standard de modélisation et

l'analyse des compositions, et a proposé d'utiliser des trans-

formations pour mapper les compositions provenant du

simplexe S^M dans \mathbb{R}^M . L'une de ces transformations est la

transformation additive par le logarithme du ratio (a^M),

définie dans Aitchison (1986, page 113), qui a été adoptée

pour la première fois dans le contexte d'une série chrono-

logique par Brunson (1987, page 75). La transformation

est donnée par $v_t = a^M(y_t) = (v_{1,t}, \dots, v_{M,t})'$, avec

$$v_{m,t} = \log \left(\frac{y_{m,t}}{y_{M+1,t}} \right), \quad m = 1, \dots, M, \quad \forall t, \quad (4)$$

où \log représente le logarithme naturel. Notons que

$y_{M+1,t} = 1 - \sum_{m=1}^M y_{m,t}$, parfois appelée valeur de remplis-

sage, est utilisée comme variable ou catégorie de référence.

La transformation inverse, appelée transformation additive

logistique, est donnée par $y_t = a_{M+1}^{-1}(v_t) = (y_{1,t}, \dots, y_{M+1,t})'$

telle que

$$y_{m,t} = \left\{ \begin{array}{l} \frac{\exp(v_{m,t})}{1 + \sum_{m=1}^M \exp(v_{m,t})} \\ \frac{1 + \sum_{m=1}^M \exp(v_{m,t})}{1} \end{array} \right. \quad m = 1, \dots, M, \quad \forall t, \quad (5)$$

La méthode de modélisation espace-état appliquée aux séries chronologiques compositionnelles ne dépend pas du choix de la variable de référence (Silva 1996) et, par conséquent, tout élément $y_{m,t} \neq y_{M+1,t}$ de y_t peut être pris comme variable de référence lorsque l'on applique la transformation additive logarithmique au vecteur des estimations d'enquête. Si les logarithmes du ratio v_t obéissent à une loi de distribution normale, la composition à $M+1$ parties est caractérisée par une distribution normale logistique additive

telles que définie dans Aitchison et Shen (1980). Pour les

séries chronologiques compositionnelles, Brunson (1987)

(VARMA) (Tiao et Box 1981) pour les séries transformées.

Nous proposons une méthode qui non seulement fournit

des prédictions et des estimations filtrées comprises entre

les bornes zéro et un et satisfait la contrainte de somme

unitaire, mais qui améliore aussi l'estimation du signal

inobservable et de ses composantes, en tenant compte de

l'erreur d'échantillonnage.

En nous inspirant de Bell et Hillmer (1990), nous

pouvons réécrire le modèle (2) sous la forme :

$$y_{m,t} = \theta_{m,t} \left(1 + \frac{\theta_{m,t}}{e_{m,t}} \right) = \theta_{m,t} u_{m,t}, \quad (6)$$

avec

$$u_{m,t} = \left(1 + \frac{\theta_{m,t}}{e_{m,t}} \right) = (1 + u_{m,t}), \quad (7)$$

où $u_{m,t} = e_{m,t}/\theta_{m,t}$ représente l'erreur d'échantillonnage

relative de la proportion estimée.

L'application de la transformation additive par le loga-

ritme du ratio définie par Aitchison (1986, page 113) au

vecteur y_t , avec les composantes données dans (2), produit

un vecteur transformé $v_t = a^M(y_t) = (v_{1,t}, \dots, v_{M,t})'$

contenu dans \mathbb{R}^M . Si nous utilisons $y_{M+1,t}$ comme variable

de référence, le vecteur transformé a pour $m^{\text{ème}}$ compo-

sante :

$$v_{m,t} = \log \left(\frac{y_{m,t}}{y_{M+1,t}} \right) = \log \left(\frac{\theta_{m,t} u_{m,t}}{\theta_{M+1,t} u_{M+1,t}} \right) = \log \left(\frac{\theta_{m,t}}{\theta_{M+1,t}} \right) + \log \left(\frac{u_{m,t}}{u_{M+1,t}} \right), \quad m = 1, \dots, M. \quad (8)$$

Partant de (8), nous pouvons écrire un modèle vectoriel

pour la série transformée sous la forme :

$$v_t = \theta_t^* + e_t^*, \quad (9)$$

avec $v_t = (v_{1,t}, \dots, v_{M,t})'$, $\theta_t^* = (\theta_{1,t}^*, \dots, \theta_{M,t}^*)'$ et

$e_t^* = (e_{1,t}^*, \dots, e_{M,t}^*)'$ où $v_{m,t} = \log(y_{m,t}/y_{M+1,t})$, $\theta_{m,t}^* = \log(\theta_{m,t}/\theta_{M+1,t})$ et $e_{m,t}^* = \log(u_{m,t}/u_{M+1,t})$, pour

$m = 1, \dots, M$. Notons que le modèle (9) a la même forme

que le modèle (1).

Pour décrire les données d'enquête, le modèle (9) doit

inclure des modèles de série chronologique pour $\{\theta_t^*\}$ ainsi

que $\{e_t^*\}$. Par conséquent, un modèle multivarié pour les

données transformées dépendra de la forme des modèles de

série chronologique utilisés pour $\{\theta_t^*\}$ et $\{e_t^*\}$.

À la section 3, nous examinons la formulation d'un

espace-état pour les données compositionnelles. À la

subséquent, Jones (1980) a procédé à une analyse primaire pour évaluer la structure du bruit d'échantillonnage, tandis que Binder et Hidiroglou (1988), Binder et Dick (1989), Pfeffermann, Burck et Ben-Tuvia (1989), Pfeffermann et Burck (1990), Pfeffermann (1991), Binder, Bleuer et Dick (1993), Pfeffermann et Bleuer (1993), Pfeffermann, Bell et Signorelli (1996), Pfeffermann, Feder et Signorelli (1998), ainsi que Harvey et Chung (2000) ont procédé à une analyse élémentaire.

L'analyse de séries chronologiques de données d'enquête nécessite aussi la modélisation du processus dont est issu le signal. Lors des tous premiers travaux, on supposait que $\{\theta_t\}$ était un processus stationnaire et que $\{y_t\}$ correspondait à la superposition de deux processus stationnaires et était donc elle-même stationnaire. Habituellement, on posait que $\{\theta_t\}$ et $\{e_t\}$ étaient des processus autoregressifs à moyennes mobiles (ARMA) et que, par conséquent, $\{y_t\}$ l'était aussi. Binder et Hidiroglou (1988) ont écrit les processus sous une forme espace-état qui a mené rapidement à l'introduction de processus non stationnaires pour le signal $\{\theta_t\}$ et, depuis, on utilise des modèles structurels comportant des tendances et des facteurs saisonniers.

L'objectif est d'améliorer l'estimation du signal inobservable et de ses composantes. Cependant, l'autocorrélation éventuelle des erreurs d'échantillonnage pourrait induire de fausses tendances que l'on confondrait avec la tendance réelle du signal, comme l'on fait remarquer Tiller (1992), ainsi que Pfeffermann, Bell et Signorelli (1996). Si l'on ne tient pas compte de la variation des erreurs d'échantillonnage, leur structure d'autocorrélation peut être absorbée dans la composante de saisonnalité ou de tendance, donc influencer sur les inférences faites d'après le modèle.

Une situation particulière qui mérite d'être examinée dans le cas des enquêtes répétées est celle où le paramètre cible unitaire $\{\theta_t\}$ est une proportion, comme le taux de chômage. La modélisation d'une série chronologique non limitée de $\{\theta_t\}$ peut produire des estimations tombant en-dehors de la fourchette $0 \leq \theta_t \leq 1$. Wallis (1987) a utilisé une transformation logarithmique pour s'assurer que les estimations soient bornées, mais a omis de tenir compte de l'erreur d'enquête. Pfeffermann (1991), Tiller (1992), Pfeffermann et Bleuer (1993), Pfeffermann, Bell et Signorelli (1996) ont ajusté des modèles espace-état aux séries de taux de chômage en tenant compte des erreurs d'enquête, mais sans recourir à la transformation logarithmique pour assurer la production d'estimations bornées.

Bien que la plupart des enquêtes soient multivariées, peu de travaux ont été consacrés à l'analyse des séries chronologiques multivariées de données d'enquête. Brunson (1987) et Brunson et Smith (1998) analysent les données multivariées provenant de sondages d'opinion en tenant compte du fait que les proportions sont bornées et comprennent une composition, mais en ne tenant pas compte de la structure des erreurs d'enquête. Ces travaux fournissent des éclaircissements utiles sur la modélisation de séries

chronologiques de proportions. Quintana et West (1988), Shepard et Harvey (1989), ainsi que Singh et Roberts (1992) ont également modélisé des données compositionnelles selon une méthode espace-état, mais ces auteurs n'ont pas abordé la question de la modélisation de la structure d'autocovariance des erreurs d'échantillonnage lorsque les compositions observées sont obtenues d'après des enquêtes répétées.

Les travaux présentés ici ont été entrepris parce que le nombre de variables étudiées par les bureaux de la statistique ont une réponse multinomiale et que l'on cherche à estimer la proportion d'unités classées dans chaque catégorie. Dans ce cas, on obtient un vecteur de proportions dont la somme des éléments est égale à l'unité et qui forme ce que l'on appelle une composition. Par conséquent, une série chronologique compositionnelle est une série chronologique multivariée comprenant des observations de compositions à chaque point dans le temps. Nous proposons d'appliquer aux séries chronologiques compositionnelles provenant d'enquêtes répétées une classe de modèles espace-état multivariés qui tiennent compte des erreurs d'échantillonnage et assurent que les estimations satisfassent les contraintes sous-jacentes imposées par les compositions. La méthode se fonde sur un modèle structurel signal plus bruit qui produit une série désaisonnalisée et des estimations de la tendance qui satisfont la contrainte de somme sous-jacente. Nous appliquons la méthode à des données compositionnelles provenant de l'Enquête sur la population active du Brésil comprenant des estimations du vecteur de proportions de situations d'activité. Nous produisons des estimations des compositions, des tendances et des séries de taux de chômage désaisonnalisées.

2. CADRE DE MODÉLISATION DES DONNÉES COMPOSITIONNELLES CHEVAUCHANTES PROVENANT D'ENQUÊTES

Nous supposons que $\{\theta_t\}$ est multivariée et que les composantes $\theta_{m,t}$ forment une composition, c'est-à-dire $0 < \theta_{m,t} < 1 \forall m, t$ et $\sum_{m=1}^{M+1} \theta_{m,t} = 1$. Dans ce cas, y_t est un vecteur d'estimations d'échantillon fondé sur les données transversales recueillies à la période t et appartient au simplexe :

$$S^M = \{y_t : 0 \leq y_{m,t} \leq 1, m = 1, \dots, M+1; \sum_{m=1}^M y_{m,t} = 1; t = 1, \dots, T\},$$

comme dans Brunson et Smith (1998). En outre, nous supposons que y_t est obtenue d'après une enquête à plan de sondage complexe, avec superposition d'unités d'un cycle à l'autre. Puisque chacune de ses composantes est sujette à

Modélisation de séries chronologiques compositionnelles d'après des données d'enquêtes répétées

D.B.N. SILVA et T.M.F. SMITH¹

RÉSUMÉ

Par séries chronologiques compositionnelles, on entend une série chronologique multivariée pour laquelle les valeurs de chaque série sont comprises entre les bornes zéro et un et la somme des séries est égale à l'unité à chaque point dans le temps. Des données présentant ces caractéristiques sont obtenues dans le cas d'enquêtes répétées, lorsque la réponse pour l'une des variables observées est multinominale, mais que l'on s'intéresse à la proportion d'unités classées dans chacune des catégories. Dans ce cas, les estimations d'enquête représentent des proportions d'un tout subordonné à une contrainte de somme unitaire. Dans le présent article, nous employons une méthode espace-état pour modéliser la série chronologique compositionnelle d'après des enquêtes répétées en tenant compte des erreurs d'échantillonnage. Nous utilisons la transformation logarithmique additive pour être certains que les prédictions et les estimations du signal soient comprises entre zéro et un et satisfassent la contrainte de somme unitaire. Nous appliquons la méthode à des données compositionnelles provenant de l'Enquête sur la population active du Brésil. Nous obtenons des estimations du vecteur des proportions et des taux de chômage. En outre, nous produisons les composantes structurelles du vecteur de signaux, tels que les événements saisonniers et les tendances.

MOTS CLÉS : Transformation logarithmique additive; séries chronologiques compositionnelles; filtre de Kalman; enquête sur la population active; enquêtes répétées; modèles espace-état.

1. INTRODUCTION

Toutes les enquêtes sont multivariées et multifonctionnelles, et la plupart sont longitudinales, avec répétition des mêmes questions au fil du temps. On distingue deux grandes catégories d'enquêtes répétées, celles avec superposition d'unités du premier degré et celles sans superposition de ces unités. Ces plans de sondage permettent tout deux une macro-analyse longitudinale des populations agrégées, mais seul le premier permet une micro-analyse de l'estimation des flux bruts ou d'une autre forme comparable de processus dynamiques de niveau unitaire. Dans le présent article, nous explorons l'analyse de séries chronologiques d'un vecteur multivarié d'agrégats de population, c'est-à-dire une macro-analyse, tout en tenant compte de l'influence des erreurs d'échantillonnage de l'enquête au moyen de données désagrégées.

Représentons par $\theta_t = (\theta_{1,t}, \dots, \theta_{M+1,t})'$ un vecteur des quantités étudiées de population au temps t et supposons que les observations sont faites à intervalles de temps égaux $t = 1, 2, \dots, T$. Représentons par $y_t = (y_{1,t}, \dots, y_{M+1,t})'$ une estimation par sondage de θ_t , fondée sur des données recueillies au temps t . Les enquêtes répétées produisent une série chronologique $\{\theta_t\}$. Si nous nous concentrons sur le vecteur inconnu de population θ_t , il est naturel d'imaginer que la connaissance de $\theta_1, \dots, \theta_{t-1}$ fournit des renseignements utiles sur θ_t , sans sous-entendre que sa valeur est parfaitement prévisible d'après $\theta_1, \dots, \theta_{t-1}$. Un moyen de

représenter cette situation consiste à considérer θ_t comme étant une variable aléatoire qui évolue stochastiquement en fonction du temps conformément à un modèle donné de série chronologique, tel que celui proposé au départ pour l'analyse univariée des données d'enquête par Bight et Scott (1973), Scott et Smith (1974) et Scott, Smith et Jones (1977). Les estimations d'enquête y_t de θ_t peuvent alors s'écrire sous la forme :

$$y_t = \theta_t + e_t \quad (1)$$

où $\{\theta_t\}$, $\{y_t\}$ et $\{e_t\}$ sont des processus aléatoires et $e_t = (e_{1,t}, \dots, e_{M+1,t})'$ représente les erreurs d'échantillonnage telles que $E(e_t | \theta_t) = 0$ et $V(e_t | \theta_t) = \Sigma_t$.

Les premiers travaux de Scott et coll. (1977) portaient sur une variable unique $\{y_t\}$ et distinguaient diverses formes des données disponibles en fonction de $\{e_t\}$. Si les seules données dont dispose l'analyste sont les estimations agrégées de population $\{y_t\}$, on parle d'analyse secondaire fondée sur une analyse secondaire des données d'enquête. Par contre, si l'on possède les enregistrements individuels de données, on peut estimer directement les variances et les covariances d'après les données et on parle alors d'analyse primaire. En outre, dans le cas d'une enquête avec renouvellement de panel, on peut se servir des estimations élémentaires (fondées sur des données recueillies auprès d'un ensemble d'unités qui s'ajoutent à l'échantillon ou le quittent toutes en même temps) pour estimer la structure de covariances des erreurs d'échantillonnage. Lors de travaux

D.B.N. SILVA, Instituto Brasileiro de Geografia e Estatística, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti 106 - Rio de Janeiro, RJ Brazil, 20231-050, courriel : denisessilva@ibge.gov.br; T.M.F. SMITH, University of Southampton, Faculty of Mathematical Studies, Highfield, Southampton, SO17 1BJ, United Kingdom, courriel : tmls@maths.soton.ac.uk.

- SMITH, T.M.F., et NIENGA, E. (1992). Méthodes robustes basées sur un modèle pour des enquêtes analytiques. *Techniques d'enquête*, 18, 201-223.
- WAND, M.P., et JONES, M.C. (1995). *Kernel Smoothing*. London : Chapman and Hall.

questions techniques. Les présents travaux ont été financés par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada.

BIBLIOGRAPHIE

BELLOUSE, D.R., et RAO, J.N.K. (2000). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, à paraître.

BELLOUSE, D.R., et STAFFORD, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

BREIDT, F.J., et OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. Soumis pour publication.

BUSKIRK, T. (1999). *Using Nonparametric Methods for Density Estimation with Complex Survey Data*. Ph.D. dissertation, Arizona State University.

EUBANK, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. New York : Marcel Dekker.

FAN, J., et GUBBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London : Chapman and Hall.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhya*, C, 37, 117-132.

GREEN, P.J., et SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London : Chapman and Hall.

HARDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press : Cambridge.

HARTLEY, H.O., et RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

HARTLEY, H.O., et RAO, J.N.K. (1969). A new estimation theory for sample surveys. II. Dans *New Developments in Survey Sampling*, (Éds. N.L. Johnson and H. Smith). New York : Wiley.

HASTIE, T.J., et TIBSHIRANI, R.J. (1990). *Generalized Additive Models*. London : Chapman and Hall.

KORN, E.L., et GRAUBARD, B.I. (1998). Scatterplots with survey data. *American Statistician*, 52, 58-69.

ONTARIO MINISTRY OF HEALTH (1992). *Ontario Health Survey : User's Guide, Volumes I et II*. Queen's Printer for Ontario.

RAO, J.N.K., et SCOTT, A.J. (1981). The analysis of categorical data from complex sample surveys : chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.

ROBERTS, G., RAO, J.N.K. et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

SÄRNDA, C.E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

SIMONOFF, J.S. (1996). *Smoothing Methods in Statistics*. New York : Springer-Verlag.

SKINNER, C.J., HOLT, D. et SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York : John Wiley and Sons.

Ici, un « lisseur » est considéré comme étant linéaire si les développements de Hastie et Tibshirani (1990), entrevoir de futurs travaux. Ce faisant, nous imitons les intégrons le contexte courant dans un cadre général laissant pourraient être utilisées. Pour décrire ces dernières, nous posons que $H = (X^T \sum_{i=1}^n X_i X_i^T)^{-1} X^T \sum_{i=1}^n Y_i$ et représentons en outre $(X^T W X)^{-1} X^T W$ par S^p . Il s'agit là de deux exemples de S . En outre, le vecteur de réponse des $\bar{y}_i = S^p y$, où y représente le vecteur de toutes les réponses d'échantillon et où S^p tient compte des poids d'échantillonage. De surcroît, le contexte habituel de régression inclut l'application d'une matrice similaire à H au vecteur complet de réponse $\hat{y}_i = H^f y$. Par conséquent, le passage de la régression ordinaire aux moyennes de régression puis au lissage par régression polynomiale locale revient à appliquer différentes matrices de lissage à y :

$$H_f y - H S^p y - S^p S^p y.$$

En général, on peut remplacer S^p par n'importe quel « lisseur » et étendre les méthodes à plusieurs covariables. Le groupement par classe des réponses présente de nombreux avantages tant théoriques que pratiques. Il est possible d'appliquer les outils de lissage classiques, comme ceux figurant dans *Splines*, sans devoir modifier le lisseur à cause de problèmes d'échantillonnage. En outre, dans le cas du modèle additif, il est possible d'invoquer le théorème de central limite applicables aux populations finies et de résoudre de la manière habituelle des questions comme le nombre de degrés de liberté, le choix du paramètre de lissage ou l'optimisation d'un critère. Dans le cas de plusieurs covariables x_1, \dots, x_q , la propriété de la dimensionnalité donne lieu à des classes peu peuplées qui ne permettent pas d'utiliser le théorème central limite. Ce problème peut être contourné de la façon habituelle en groupant les résidus partiels par classe, une dimension à la fois. Dans ce cas, on utiliserait les lisseurs $S_j S^{b_j}$ $j = 1, \dots, q$ dans un algorithme de rétro-aljustement.

Nous nous proposons d'étudier les modèles additifs et les modèles additifs généralisés de la façon susmentionnée et d'appliquer ces méthodes à l'analyse des données d'enquête complexe.

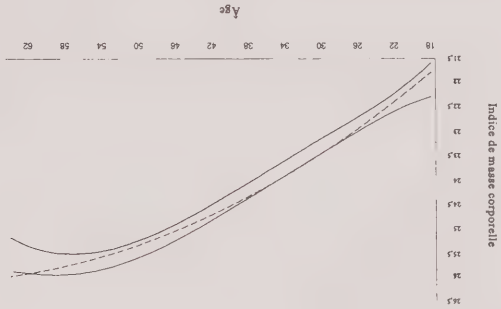
REMERCIEMENTS

Les auteurs remercient Rob Tibshirani de ces commentaires judicieux au sujet du présent article et les examinateurs, pour leurs remarques qui leur ont permis d'améliorer la présentation de l'article et de préciser certaines

d'échantillonnage due à la taille plus petite des échantillons aux extrémités. La figure 5 donne à penser que l'IMC augmente légèrement à mesure que la consommation de matières grasses augmente. Comme on disposait du fichier complet de données d'enquête, les courbes de régression ont pu être produites pour toutes les variables en se servant de SUDAAN.

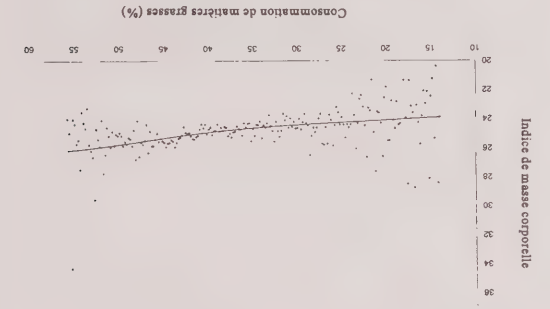
À la figure 4, les courbes en trait plein présentent les bandes de confiance à 95 % calculées d'après (6) et la courbe en pointillé correspond à la courbe de régression paramétrique polynômiale de second degré. Comme la courbe en pointillé s'approche de la limite pour les femmes dans la trentaine et tombe en dehors des bandes pour celles au début de la soixantaine, une équation polynômiale de second degré ne décrit qu'à peine adéquatement la relation entre l'IMC et l'âge. Le choix d'un autre modèle serait préférable. La figure 6 montre les mêmes bandes de confiance à 95 % pour la consommation de matières grasses exprimée en pourcentage de la consommation totale d'énergie. Dans ce cas, la courbe en pointillé correspond à la droite de régression linéaire simple de l'IMC sur la consommation de matières grasses. Dans le cas de la consommation de matières grasses, la courbe se situe entièrement dans les bandes de confiance, si bien que la régression linéaire simple semble être un modèle descriptif adéquat de la relation.

Figure 4. Bandes de confiance pour la tendance de l'IMC



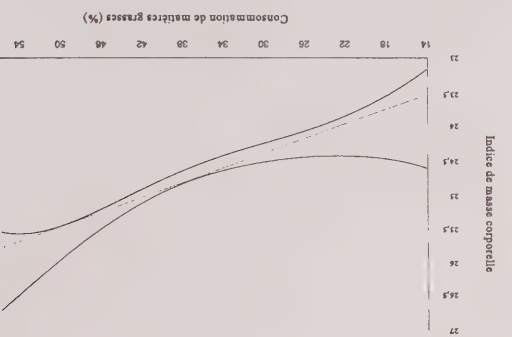
Si les données ont été groupées par classe dont la largeur correspond à la précision des données, comme dans le cas susmentionné de l'âge et que l'analyse exploratoire est complète, on peut s'arrêter. En effet, les estimations et les estimations de leur variance ainsi obtenues sont égales respectivement aux estimations et aux estimations de leur variance calculées d'après les données brutes. Pour le membre de (3) peut être exprimé sous forme de somme sur l'échantillon des poids d'échantillonnage multiplié par une nouvelle mesure obtenue d'après la mesure brute de y multipliée par une valeur appropriée extraite de $e^T(X_T^T W^X X^X)^{-1} X_T^T W^X$ correspond à W^X dont on a supprimé les p_i . Ces mesures ajustées de y peuvent être

Figure 5. Tendance de l'IMC en fonction de la consommation de matières grasses



5. ORIENTATIONS FUTURES

Figure 6. Bandes de confiance pour la tendance de l'IMC en fonction de la consommation de matières grasses



entrées dans le programme SUDAAN ou STATA pour obtenir l'estimation requise de la variance approximative. Il se pourrait que le groupement par classe n'ait pas eu la même finesse que la précision des données ou que l'on ait laissé tomber certaines classes aux queues de la distribution de x en raison de la paucité des données. Ces situations se sont produites l'une et l'autre lors de l'analyse de la relation entre l'IMC et la consommation de matières grasses. Une fois l'analyse exploratoire achevée, nous pouvons revenir à un modèle et un paramètre de lissage finals si nous adoptons une méthode non paramétrique lors de l'analyse finale et appliquer le modèle aux données brutes, en obtenant les estimations de la variance au moyen du programme SUDAAN ou STATA, au besoin. Selon le manque de finesse du groupement par classe et le nombre de classes supprimées en raison de la paucité des données, les estimations de la variance obtenues d'après les données brutes seront approximativement les mêmes que celles calculées d'après les données groupées par classe.

À l'instar de Bellhouse et Stafford (1999), nous adaptons ici une méthode moderne de lissage à l'analyse des données d'enquête complexe. La méthode exposée n'est qu'un

appropriée comprises dans ces bandes représenterait alors une modélisation raisonnable des données. Nous obtenons les bandes de régression polynomiale $100(1 - \alpha)\%$ en portant en graphique

$$(6) \quad \hat{m}(x) \pm z_{\alpha/2} \sqrt{\hat{V}^d(m(x))}$$

sur une fourchette de valeur de x , où $z_{\alpha/2}$ représente le $100(1 - \alpha/2)$ percentile de la distribution normale type, où $\hat{m}(x)$ est déterminé d'après (3) et où $\hat{V}^d(m(x))$ correspond à l'équation (5) dans laquelle \mathbf{V} est remplacé par son estimation d'échantillon $\hat{\mathbf{V}}$.

Nous pouvons obtenir la droite de régression paramétrique à tester de deux façons différentes, suivant les données d'échantillon dont on dispose. Si nous possédons le fichier complet de données d'échantillon, y compris les poids d'échantillonnage, nous pouvons utiliser la méthode de régression type figurant, par exemple, dans le logiciel SUDAAN. Si nous ne disposons que des données groupées par classe, plus précisément les estimations d'enquête \hat{y}_i , ainsi que la matrice estimative variance-covariance $\hat{\mathbf{V}}$, nous devons utiliser une autre méthode.

Pour cette deuxième méthode, supposons que $\mathbf{m}(x) = \mathbf{x}_i^T \boldsymbol{\beta}$, où $\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, \dots, x_{iq})^T$ et où $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_q)$ représente le vecteur des coefficients de régression. En ce qui concerne la population finie, nous supposons que $\hat{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, où le terme d'erreur correspond à l'écart entre la valeur réelle pour la population finie et la valeur du modèle. Par souci de simplicité, supposons que la moyenne des erreurs est 0 et que la matrice de variance-covariance est $\sigma^2 \mathbf{I}$. Puisque les données correspondent aux estimations d'enquête \hat{y}_i avec la matrice de variance-covariance $\hat{\mathbf{V}}$, le modèle opérationnel est

$$(7) \quad \hat{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \delta_i$$

où la moyenne des δ_i est 0 et la matrice de variance-covariance, $\Sigma = \sigma^2 \mathbf{I} + \mathbf{V}$. L'estimation ordinaire de $\boldsymbol{\beta}$ par les moindres carrés pondérés est donnée par

$$(8) \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Sigma^{-1} \hat{\mathbf{y}}$$

où la i ème ligne de \mathbf{X} est \mathbf{x}_i^T , $i = 1, \dots, k$. En ce qui concerne l'analyse des données, il est nécessaire de remplacer Σ dans (8) par son estimation $\hat{\Sigma}$. Comme l'estimation d'enquête de \mathbf{V} est $\hat{\mathbf{V}}$, il ne reste qu'à obtenir une estimation de σ^2 . Nous pouvons, pour cela, utiliser $\text{rss} = (\hat{\mathbf{y}} - \mathbf{X} \hat{\boldsymbol{\beta}})^T (\hat{\mathbf{y}} - \mathbf{X} \hat{\boldsymbol{\beta}})$, c'est-à-dire la somme des carrés des résidus, de deux façons.

La première méthode consiste à produire une approximation de la somme des résidus attendue compte tenu du modèle (7) et de calculer directement σ^2 . Si nous utilisons le développement $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \dots$, nous obtenons

$$(9) \quad E(\text{rss}) = (n - q - 1)\sigma^2 + \text{tr}(\mathbf{V}) - \text{tr}(\mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}).$$

Pour estimer σ^2 , nous établissons l'égalité entre rss et le deuxième membre de l'équation (8), où \mathbf{V} est remplacé par $\hat{\mathbf{V}}$, puis nous calculons la valeur de σ^2 . Cette façon de procéder mène à une méthode itérative d'ajustement du modèle. Nous obtenons une première estimation de $\boldsymbol{\beta}$ à partir de (8) où \mathbf{V} est remplacé par l'estimation d'enquête $\hat{\mathbf{V}}$. Puis, nous estimons σ^2 au moyen de (9) et nous obtenons une nouvelle estimation de $\boldsymbol{\beta}$ en utilisant $\hat{\Sigma} = \hat{\sigma}^2 \mathbf{I} + \hat{\mathbf{V}}$. Le procédé est répété jusqu'à la convergence de l'estimation de σ^2 . Si l'estimation de σ^2 est négative, nous fixons sa valeur à 0. La deuxième méthode d'estimation de σ^2 consiste, en premier lieu, à traiter dans l'équation (7) les erreurs comme des variables normales multivariées. Puis, nous pouvons obtenir un profil de vraisemblance de σ^2 en remplaçant $\boldsymbol{\beta}$ et \mathbf{V} par leurs estimations. Le terme le plus influent dans ce profil de vraisemblance est

$$(10) \quad \mathbf{r}^T (\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \mathbf{r},$$

où $\mathbf{r} = \hat{\mathbf{y}} - \mathbf{X}(\mathbf{X}^T (\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 \mathbf{I} + \hat{\mathbf{V}})^{-1} \hat{\mathbf{y}}$ représente le vecteur des résidus. Une approximation de l'estimation qui minimise l'expression (10).

Pour illustrer le problème de l'adéquation de la régression paramétrique, nous examinons deux variables distinctes de l'Enquête sur la santé en Ontario et leur lien à l'indice de masse corporelle (IMC). Ces deux variables sont l'âge et la consommation de matières grasses exprimée en pourcentage de la consommation totale d'énergie. Pour l'âge, le groupement par classe se fait naturellement avec une largeur de fenêtre correspondant à la précision des données enregistrées. Nous avons limité les valeurs de la variable d'âge à la fourchette de 18 à 65 ans puisque l'IMC n'est pas applicable en dehors de celle-ci et nous avons exprimé l'âge en années. La figure 1 donne le nuage de points obtenus lorsque l'on porte en graphique la valeur de l'IMC en fonction de l'âge ainsi que la courbe de régression polynomiale locale correspondante. Les données d'enquête sur la consommation de matières grasses exprimée en pourcentage ont été enregistrées à trois décimales près. Etant donné la grande dispersion des données aux extrémités du domaine, nous examinons la consommation de matières grasses pour la fourchette de 14 % à 56 % de la consommation totale d'énergie. En outre, nous avons groupé les données par classe associée à la covariable (consommation de matières grasses) en utilisant les classes 14,0 à 14,2, 14,2 à 14,4, et ainsi de suite; nous avons utilisé comme valeur de x_i les points milieu des classes (14,1, 14,3 et ainsi de suite). Puis, nous avons calculé l'estimation d'échantillon \hat{y}_i de l'IMC pour chaque classe. Ce sont les données groupées par classe qui figurent à la figure 5 dans le nuage de points de l'IMC en fonction de la consommation de matières grasses. Dans cette figure, la courbe en trait plein est la courbe de régression polynomiale locale où $q = 1$ pour l'IMC en fonction de la consommation de matières grasses. Comme à la figure 3, la plus forte variabilité aux extrémités reflète la plus grande variabilité

de constater que l'IMC augmente de façon plus ou moins linéaire avec l'âge jusqu'à environ 50 ans. Au début de la cinquantaine, sa valeur augmente moins rapidement, passe par un sommet à environ 55 ans, puis commence à diminuer. Si l'on ne représente graphiquement que les courbes de tendance de l'IMC et de l'IMCD observées pour les femmes, telles qu'illustrées à la figure 2, on constate qu'en moyenne, à tout âge, les femmes souhaitent réduire leur IMC d'environ deux unités.

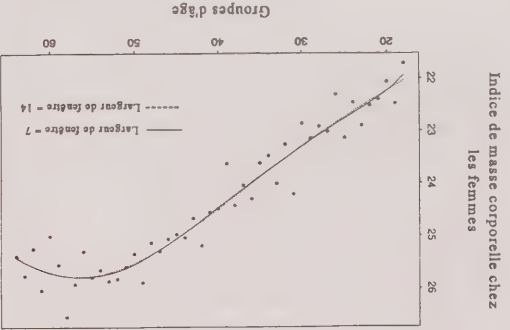


Figure 1. Tendance de l'IMC selon l'âge chez les femmes

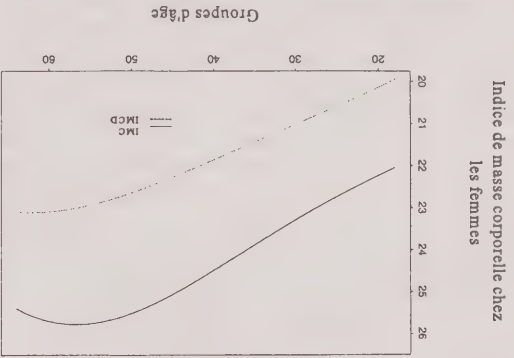


Figure 2. Tendances selon l'âge chez les femmes

Dans d'autres situations, il est commode de construire sur x des classes dont la largeur est supérieure à la précision des données. Pour étudier la relation entre le poids que les femmes désirent ($IMCD = \hat{y}_i$) et leur poids réel ($IMC = x_i$), nous avons regroupé les valeurs de x . Puisque les données étaient fort peu nombreuses pour les valeurs de l'IMC inférieures à 15 ou supérieures à 42, nous n'en n'avons pas tenu compte dans l'analyse. Les autres groupes étaient 15,0 à 15,2, 15,3 à 15,4 et ainsi de suite, où la valeur de x_i est choisie comme étant la valeur moyenne dans chaque groupe. Nous avons fait le groupement par classe de cette façon en guise d'illustration afin d'obtenir une grande gamme de classes non vides, espacées de façon égale. Pour chaque groupe, nous avons calculé l'estimation

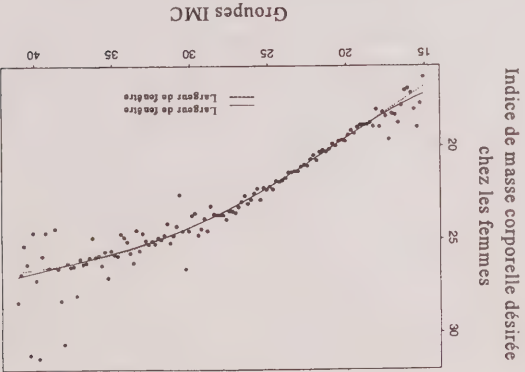


Figure 3. Tendance de l'IMC en fonction de l'IMCD

4. RÉGRESSION PARAMÉTRIQUE CONTRE NON PARAMÉTRIQUE

La régression polynomiale locale nous permet d'obtenir de façon non paramétrique une relation fonctionnelle entre y et x . Toutefois, le choix d'un modèle paramétrique serait raisonnable également. Par exemple, si nous examinons la figure 1 donnant l'indice de masse corporelle en fonction de l'âge, nous pourrions envisager le modèle paramétrique représentant une relation quadratique entre y et x . Nous pourrions aussi vouloir vérifier, dans la figure 2, si les deux droites sont parallèles ou, de façon équivalente, si l'écart entre l'indice de masse corporelle et l'indice de masse corporelle désirée est invariable selon l'âge chez les femmes. Il faudrait pour cela modéliser les courbes de tendances sous forme de polynômes de deuxième degré et vérifier si les paramètres associés au terme linéaire, sont les mêmes que ceux associés au terme quadratique, de même que ceux associés au terme linéaire, sont les mêmes pour les deux courbes de tendance. En tous cas, la question qui se pose est celle de savoir s'il est possible de modéliser adéquatement les données au moyen d'une relation polynomiale entre y et x . L'une des méthodes que nous proposons pour répondre à cette question consiste à calculer les bandes de confiance en s'appuyant sur la régression polynomiale locale. Nous pouvons imaginer ces bandes comme définissant une région où les représentations du modèle sont acceptables. Une courbe de régression paramétrique

3. EXEMPLES FONDÉS SUR LES DONNÉES DE L'ENQUÊTE SUR LA SANTÉ EN ONTARIO

Nous illustrons les méthodes de régression polynomiale locale au moyen des données de l'Enquête sur la santé en Ontario (ministère de la Santé de l'Ontario 1992). Cette enquête a été réalisée en 1990 auprès d'un échantillon en grappes stratifié à deux degrés. L'objectif était d'évaluer l'état de santé des résidents de l'Ontario et de recueillir des données sur les facteurs de risque associés aux causes principales de morbidité et de mortalité dans la province. L'enquête a été conçue de sorte que les données soient compatibles avec celles de l'Enquête Santé Canada réalisée en 1978-1979. En tout, un échantillon de 61 239 personnes a été sélectionné parmi la population relevant de 43 bureaux de santé de l'Ontario. Les bureaux de santé, qui représentent les strates de base, ont été subdivisés chacun en une strate rurale et une strate urbaine, si bien que l'on a obtenu en tout 86 strates. Les unités de premier degré dans les strates étaient les secteurs de dénombrement tels que définis pour le Recensement du Canada de 1986. En moyenne, 46 secteurs de dénombrement ont été choisis dans chaque strate. Puis, des logements ont été sélectionnés dans chaque secteur de dénombrement, au nombre d'environ 15 pour les secteurs de dénombrement urbains et de 20 pour les secteurs de dénombrement ruraux. Enfin, des renseignements ont été recueillis sur tous les membres des ménages vivants dans les logements sélectionnés.

Plusieurs caractéristiques de l'état de santé ont été évaluées. Nous nous concentrons ici sur l'une des variables continues de l'enquête, à savoir l'indice de masse corporelle (IMC). Pour calculer ce dernier, qui donne une évaluation du poids, on divise le poids exprimé en kilogrammes par le carré de la taille exprimée en mètres. L'indice ne s'applique pas aux adolescents ni aux adultes de plus de 65 ans ni aux femmes enceintes ou qui allaient. L'indice s'étend de 7,0 à 45,0. Un IMC inférieur à 20,0 est souvent associé à des problèmes de santé tels que les troubles des comportements alimentaires. Une valeur supérieure à 27,0, quant à elle, est associée à des problèmes de santé tels que l'hypertension et la maladie coronarienne. L'IMC est relié à une autre mesure, l'indice de masse corporelle désirée (IMCD). L'IMCD se calcule de la même façon que l'IMC en remplaçant le poids réel par le poids désiré. En tout, 44 457 réponses ont été recueillies pour l'IMC et 41 939 pour l'IMCD.

Lorsqu'il n'existe que quelques réalisations distinctes de x , le groupement par classe en fonction de x se fait naturellement. Par exemple, si l'on étudie la relation entre l'indice de masse corporelle (IMC) et l'âge, les valeurs déclarées de l'âge du répondant correspondent uniquement à des nombres entiers. Dans la figure 1, les points pleins sont les estimations d'enquête par domaine de l'IMC moyen (\bar{y}_i) pour les femmes, pour chaque année d'âge de 18 à 65 ans (x_i). Les courbes en trait plein et en pointillé représentent le graphe de $\hat{m}(x)$ en fonction de x pour les largeurs de

L'estimation $\hat{m}(x)$ et ses deux premiers moments finie pour les valeurs distincts de x et par \bar{y} , le vecteur des $\bar{y} = (\bar{y}_1, \dots, \bar{y}_K)^T$ le vecteur des moyennes de la population nous adoptons la notation. Représentons par exemple, dans Wand et Jones (1995, chapitre 5.3) dont sont exactement les mêmes que celles qui figurent, par peuvent être exprimées en notation matricielle. Les formes

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 - x & \dots & (x_1 - x)^p \\ 1 & x_2 - x & \dots & (x_2 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_K - x & \dots & (x_K - x)^p \end{bmatrix}$$
$$\mathbf{W} = \frac{1}{h} \text{diag} \left(p_1 K(x_1 - x)/h, \dots, p_K K(x_K - x)/h \right),$$
$$p_K K(x_K - x)/h, \dots, p_1 K(x_1 - x)/h \Big) \Big) \Big)$$

La matrice \mathbf{W} est \mathbf{W}_x où p est remplacé par p . Alors,

$$\hat{m}(x) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \bar{\mathbf{y}}, \quad (3)$$

où \mathbf{e} est le vecteur $k \times 1$, soit $(1, 0, 0, \dots, 0)^T$. L'espérance approximative, fondée sur le plan de sondage, de $\hat{m}(x)$ est où E_p représente l'espérance compte tenu du plan de

$$E_p(\hat{m}(x)) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \bar{\mathbf{y}}, \quad (4)$$

Nous pouvons également considérer (4) comme une estimation lissée de $m(x)$, de sorte que $\hat{m}(x)$ est également une estimation de $m(x)$. En établissant (4), nous notons que $E_p(\bar{\mathbf{y}}) = \bar{\mathbf{y}}$ et $E_p(\mathbf{W}_x) = \mathbf{W}_x$ pour une grande taille d'échantillon n . En outre, dans (3) nous pouvons écrire $\mathbf{W}_x = \mathbf{W}_x^* + \mathbf{A}$, où $\mathbf{A} = \mathbf{W}_x^* - \mathbf{W}_x$. Nous utilisons les deux premiers termes dans le développement $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \dots$ en tant qu'approximation pour achever la dérivation. Selon les mêmes techniques, nous obtenons pour la variance approximative fondée sur le plan de sondage l'expression

$$V_p(\hat{m}(x)) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{V} \mathbf{W}_x \mathbf{X}_x (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{e}. \quad (5)$$

Les équations (4) et (5) ont été obtenues en ignorant les termes d'ordre plus élevé $1/n$. Nous obtenons une estimation de la variance $V_p(\hat{m}(x))$ est substituant l'estimation d'enquête $\hat{\mathbf{V}}$ à \mathbf{V} et $\hat{\mathbf{W}}_x$ à \mathbf{W}_x dans (5).

Le groupement des données par classe en fonction de la covariable x présente plusieurs avantages pour l'analyse exploratoire des données.

- Dans le cas de grandes enquêtes, un graphe de \hat{y}_i en fonction de x_i pourrait fournir plus de renseignements et être moins encombré que le graphe des données brutes.
- En invoquant un théorème central limite applicable à la population finie pour \hat{y}_i et en émettant une hypothèse de superpopulation pour \hat{y}_i , nous pouvons établir un modèle hypothétique relativement simple pour \hat{y}_i si bien que l'analyse pourra facilement se concentrer sur le problème essentiel considéré ici, à savoir la détermination de la fonction de tendance dans x .
- Une fois l'estimation \hat{V} obtenue, il est facile de réaliser une gamme étendue d'analyses exploratoires puissantes des données dans des langages tels que S-Plus. En revanche, si l'on analyse les données brutes, il faut continuer de se servir de SUDAN ou de STATA pour produire les estimations appropriées de la variance.
- Le groupement des données par classe permet d'adopter une méthode d'analyse par régression qui fait pendant à d'autres méthodes non paramétriques d'analyse des données d'enquête. Par exemple, dans le cas de l'analyse catégorique des données réalisée au départ par Rao et Scott (1981), de la méthode de régression logistique de Roberts, Rao et Kumar (1987) ou du modèle linéaire généralisé de Beilhouse et Rao (2000), les statistiques et les distributions associées sont obtenues par calcul des estimations par sondage des moyennes ou des proportions par domaine.

Dans le cas de la superpopulation, nous supposons que le modèle est tel que $E_m(\hat{y}_i) = m(x_i)$, où E_m représente l'espérance de la superpopulation. Nous supposons en outre que, lorsque nous passons à une suite continue de valeurs sur x , $m(x)$ devient une fonction lisse. La fonction $m(x)$ est celle à laquelle nous nous intéressons, en dernière analyse, pour l'estimation. À la section 2, nous présentons les méthodes de régression polynormale utilisées pour estimer $m(x)$. À la section 3, nous appliquons ces méthodes aux données de l'enquête sur la santé en Ontario de 1990. À la section 4, nous déterminons si les méthodes classiques de régression polynormale auraient donné d'aussi bons résultats pour la modélisation de $m(x)$. Enfin, à la section 5, nous examinons certaines orientations futures de ces travaux. En général, nous adoptons la notation de Wand et Jones (1995) pour discuter de la régression polynormale locale.

2. MÉTHODOLOGIE DE BASE

Pour la régression polynormale locale, on obtient l'estimation de $m(x)$ pour toute valeur de x par minimisation de

$$(1) \quad -\beta^q(x_i - x)^q \int_2^2 K(x_i - x)/h \quad \text{par rapport à } \beta_0, \beta_1, \dots, \beta^q. \text{ Les valeurs qui minimisent l'équation (1) sont représentées par } \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}^q. \text{ En outre, pour les valeurs données de } x, \hat{m}(x) = \hat{\beta}_0. \text{ Dans (1), le noyau } K(t) \text{ est une fonction symétrique pour laquelle } \int K(t) dt = 1, \int t K(t) dt = 0, 0 < \int t^2 K(t) dt < \infty \text{ et } \int_2^2 [K(t)]^2 dt < \infty. \quad (2)$$

En outre, dans (1), h représente la largeur de fenêtre du noyau. Lors de la minimisation de (1) pour estimer les paramètres de la régression polynormale locale, le groupement par classe en fonction de x peut se faire de deux façons. La première consiste à choisir pour dimension des classes la précision des données enregistrées, si bien que \hat{y}_i est calculé pour chaque réalisation distincte de x . Dans d'autres situations, il peut être pratique de procéder à un groupement par classe en fonction de x plus grossier que celui fondé sur la précision des données.

Si nous passons de l'échantillon à la population, nous maintenons la même largeur de fenêtre h . Cette démarche est différente de celles de Breidt et Opsomer (2000) et de Baskirk (1999) qui émettent l'hypothèse d'un paramètre de lissage h_N pour le lissage au niveau de la population finie et complète. Ici, cette démarche produirait une fonction $m_N(x)$, c'est-à-dire la version lissée de \hat{y}_i pour la population finie si l'on utilise le paramètre de lissage h_N , en temps que paramètre étudié de la population finie suivi par $m(x)$, la fonction lissée hypothétique dans les conditions asymptotiques. Ici, nous maintenons la valeur de h constante étant donné la façon dont a été fait le groupement par classes; la structure des classes est la même pour l'échantillon que pour la population. Le choix du paramètre de lissage h dépend de l'intervalle entre les valeurs de x et de la variabilité des données (Green et Silverman 1994, pages 43-44). L'intervalle entre les valeurs de la covariable est habituellement un déterminant dominant de la valeur de h . Puisque l'intervalle ne varie pas lorsque l'on passe de l'échantillon à la population finie, et qu'il change uniquement lorsque l'on applique les conditions asymptotiques, nous maintenons $h_N = h$.

Korn et Graubard (1998) proposent une fonction objective légèrement différente de celle représentée par l'équation (1). Ils remplacent dans cette dernière la somme sur les classes par la somme sur toutes les unités échantillonnées et p_i par les poids d'échantillonnage. La fonction objective de Korn et Graubard s se réduit à (1) plus un terme comprenant la somme pondérée des carrés des écarts des observations de l'échantillon par rapport aux moyennes calculées par classe où les poids sont leur somme soit égale à l'annulation des écarts. Par conséquent, l'estimation de $m(x)$ est la même dans les deux cas.

Régression polynômiale locale dans le cas des enquêtes complexes

D.R. BELLHOUSE et J.E. STAFFORD¹

RÉSUMÉ

Certaines méthodes de régression polynômiale locale sont présentées pour faciliter l'analyse exploratoire des données provenant d'enquêtes à grande échelle. Les méthodes proposées s'appuient sur le groupement des données par classe (*binning*) sur la variable x , ainsi que sur le calcul des estimations d'enquêtes pertinentes de la moyenne des valeurs de y dans chaque classe (ou fenêtre). Si le groupement par classe sur x est exécuté en prenant la précision des données enregistrée comme largeur de fenêtre, la méthode revient à appliquer les poids de sondage au critère standard utilisé pour obtenir des estimations par régression polynômiale locale. On considère aussi l'autre solution qui consiste à procéder à la régression polynômiale classique et on propose un critère pour décider si la méthode non paramétrique est ou non préférable à la méthode classique de modélisation. Des exemples tirés de l'Enquête sur la santé en Ontario de 1990 sont donnés à titre d'illustration.

MOTS CLÉS : Covariables; analyse exploratoire des données; lissage par la méthode du noyau; régression.

1. INTRODUCTION

Suite aux travaux de Fuller (1975), les techniques de régression linéaire multiple ont été étudiées et utilisées à grande échelle dans le contexte des enquêtes par sondage. Au moins trois chapitres de l'ouvrage de Skinner, Holt et Smith (1989) sont consacrés à ce sujet. Ici, nous nous limitons au cas où il n'existe qu'une seule covariable x pour la variable étudiée y , si bien que nous pouvons considérer la régression polynômiale, ainsi que la régression linéaire simple. Dans ce contexte, nous pourrions aussi envisager l'approche non paramétrique de la régression polynômiale locale, qui, dans le cas de variables aléatoires indépendantes et distribuées de façon identique, est décrite par Hardle (1990), Wand et Jones (1995), Fan et Gijbels (1996), Simonoff (1996) et Eubank (1999). Korn et Graubard (1998) ont introduit, en se servant des poids de sondage, l'utilisation de la régression polynômiale locale pour l'affichage graphique des données d'enquête complexe. Cependant, ils n'ont donné aucune propriété statistique de leurs méthodes. Smith et Njenga (1992) ont appliqué des techniques de lissage par régression, selon la méthode du noyau, pour obtenir des estimations robustes de la moyenne et des paramètres de régression dans le cas d'un modèle hypothétique de superpopulation. Ici, nous utilisons la régression polynômiale locale comme outil d'exploration en vue de découvrir la relation entre y et sa covariable x . Nous supposons que la covariable x est mesurée sur une échelle continue. Étant donné la précision à laquelle les données sont enregistrées dans le fichier d'enquête et la taille de l'échantillon, il existera plusieurs observations de y pour nombre de valeurs distinctes de x . Harley et Rao (1968, 1969) ont exploité cette caractéristique des données d'enquête à grande échelle dans leur méthode *scale-load*

d'estimation des paramètres d'une population finie. Ici, nous exploitons cette même caractéristique des données pour examiner la relation entre y et sa covariable x . Reconnaisant que les données pourraient être groupées naturellement par classe de largeur correspondant à la précision des données, nous pouvons faire un pas de plus et construire des classes ou fenêtres de plus grande dimension. Dans ce contexte, nous examinons l'effet du plan d'échantillonnage sur les estimations et sur les moments de deuxième ordre. Supposons que, dans une population finie de taille N , x prend k valeurs distinctes, si bien qu'un groupement naturel par classe a eu lieu, ou que x a été catégorisée en k classes dont la largeur est supérieure à la précision des données. Soit x_i la valeur de x représentant la $i^{\text{ème}}$ classe, et supposons que l'intervalle entre les valeurs x_i est constant. L'intervalle ou largeur de la classe est égal à $b = x_i - x_{i-1}$. La moyenne des valeurs de y dans la population finie associée à x_i est \bar{y}_i . Nous supposons que la structure d'un échantillon de taille n tiré de cette population est la même que celle de la population en ce sens qu'il compte k classes. D'après les données d'échantillon, nous calculons l'estimation par p_i . Représentons par p_i la proportion d'observations associées à la valeur x_i dans la population finie. L'estimation d'échantillon de cette proportion est \hat{p}_i . Nous supposons que \bar{y}_i et \hat{p}_i sont asymptomatiquement non biaisées, au sens de Särndal, Swensson et Wretman (1992, pages 166-167), pour \bar{y}_i et \hat{p}_i , respectivement. Les estimations d'échantillon \bar{y}_i pour $i = 1, \dots, k$ sont caractérisées par la matrice de variance-covariance V . Si l'on considère les valeurs distinctes de x_i comme des domaines, on peut obtenir facilement la matrice estimative de variance-covariance V en se servant de logiciels comme SUDAAN ou STATA.

¹ D.R. Bellhouse Department of Statistical and Actuarial Sciences, Western Science Centre, University of Western Ontario, London, Ontario N6A 5B7, courriel : bellhouse@stats.uwo.ca; J.E. Stafford, Department of Public Health Sciences, Faculty of Medicine, McMaster Building, University of Toronto, Toronto, Ontario, M5S 1A8, courriel : stafford@uistat.toronto.edu.

SHAO, J., et RAO, J.N.K. (1994). Standard errors for low income proportions estimated from stratified multi-stage samples *Sankhyā B*, Special Volume 55, 393-414.

SHAO, J., et SITTER, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

SHAO, J., et WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics*, 20, 1571-1593.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

SITTER, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

SHAO, J., CHEN, Y. Et CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

RAO, J.N.K., et WU, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

RUBIN, D.B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

RUBIN, D.B., et SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Étant donné les résultats du tableau 3, nous ne considérons l'instabilité relative, IR, que pour v_B et v_{BRR} . Nous limitons également notre présentation à $p = 0,3$ et $p = 0,6$ puisque les résultats sont qualitativement les mêmes dans les trois autres cas. Les résultats sont présentés au tableau 4. On constate, quoique les différences soient faibles, que v_B est légèrement plus stable que v_{BRR} . Cette différence s'observe en général pour toutes les valeurs de p et de p . Nous avons également inclus dans les simulations l'estimateur jackknife corrigé de Rao et Shao (1992) et l'estimateur BRR corrigé de Shao, Chen et Chen (1998), pour $\theta = Y$ et de nouveau, v_B est uniformément plus stable. Par exemple, pour $p = 0,3$ et $p = 0,6$ dans le tableau 4, l'IR est égale à 0,27 pour l'estimateur jackknife corrigé ainsi que pour l'estimateur BRR corrigé. Cette situation pourrait tenir au fait que la méthode de réimputation estime mieux la composante de la variance due à l'imputation que la méthode corrigée, à condition que la taille du nouvel échantillon soit suffisamment grande pour éliminer l'erreur de l'approximation de Monte Carlo, comme cela est le cas dans nos simulations. Toutefois, si le nombre de réimputations est moyen (comme dans les répliques équilibrées avec réimputation ou l'estimation bootstrap avec $B = 1\ 000$), cet avantage n'est pas entièrement réalisé.

Tableau 4

IR pour v_B et v_{BRR} avec $p = 0,3$ et $p = 0,6$

Paramètre estimé	v_{BRR}	v_B	v_{BRR}	v_B	Imputation aléatoire	Imputation aléatoire corrigée
$F(t) = 0,0625$	0,27	0,23	0,59	0,57	0,56	0,35
$F(t) = 0,2500$	0,35	0,32	0,37	0,28	0,26	0,28
$F(t) = 0,5000$	0,27	0,23	0,26	0,30	0,28	0,28
$F(t) = 0,7500$	0,29	0,26	0,46	0,48	0,46	0,46
$F(t) = 0,9375$	0,48	0,46	0,48	0,48	0,46	0,46

7. CONCLUSION

Nous proposons une méthode bootstrap à demi-échantillon répété et une méthode par répliques équilibrées d'estimation de la variance en cas d'imputation aléatoire qui tiennent compte de la variance due à l'imputation grâce à une réimputation lors de chaque répétition, selon la même méthode d'imputation aléatoire que celle utilisée pour l'échantillon original. Les méthodes à demi-échantillon répété sont valides en cas d'échantillonnage stratifié à plusieurs degrés, même si le nombre d'UPÉ échantillonnées dans chaque strate est très faible, par exemple, 2. L'élément clé de ces méthodes est que la taille du rééchantillon de strate est égale à celle de l'échantillon original sans que l'on recoure au rééchantonnement. Nous obtenons ainsi une méthode unifiée, applicable

REMERCIEMENTS

Les travaux de Hiroshi Saigo ont été financés par des bourses de la Promotion and Mutual Aid Corporation for Private Universities of Japan et la Japan Economic Research Foundation. Ceux de Jun Shao ont été financés par la National Science Foundation Grant DMS-0102223, et la National Security Agency Grant MDA904-99-1-0032. Les travaux de Randy R. Sitter ont été financés par une bourse du Conseil de recherches en sciences naturelles et en génie du Canada. Les auteurs remercient tous les examinateurs de leurs commentaires constructifs.

BIBLIOGRAPHIE

CHEN, J., RAO, J.N.K. et SITTER, R.R. (2000). Adjusted imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

EFFRON, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89, 463-479.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

MANTEL, H.J., et SINGH, A.C. (1991). Standard errors of estimates of low proportions: A proposed methodology. Rapport technique, Statistique Canada.

MCCARTHY, P.J. (1969). Pseudoreplication half samples. Review of the International Statistical Institute, 37, 239-264.

NIGAM, A.K., et RAO, J.N.K. (1996). On balanced bootstrap, for stratified multistage samples. *Statistica Sinica*, 6, 199-214.

RAO, J.N.K. (1993). Linearization variance estimators under imputation for missing data. Rapport technique, Laboratory for Research in Statistics and Probability, Carleton University.

RAO, J.N.K., et SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., et SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

RAO, J.N.K., et WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for non-linear statistics. *Journal of the American Statistical Association*, 80, 620-630.

Tableau 3
% BR pour v_B, v_{B2}, v_{BRR} et v_{BRR2}

Paramètre estimé	v_{BRR}	v_{BRR2}	v_B	v_{B2}	v_{BRR}	v_{BRR2}	v_B	v_{B2}	Imputation aléatoire	
Y	0,00	21,54	0,79	21,60	0,46	19,64	1,24	19,51	p = 0,1 et p = 0,6	
	-1,09	15,92	-0,52	15,88	0,85	14,86	1,80	15,08		
	-0,13	19,44	0,62	19,55	0,55	10,73	1,24	10,76		
	-0,36	21,68	0,52	21,55	-0,36	10,98	0,54	11,31		
	-0,84	19,89	0,13	20,09	0,81	19,12	1,39	18,91		
Y	-0,63	15,06	0,36	15,37	-1,65	10,97	-1,08	11,13	p = 0,1 et p = 0,8	
	-1,99	10,30	-1,72	10,16	-1,65	10,97	-1,08	11,13		
	-1,27	13,65	-0,88	13,30	-0,95	8,89	-0,52	8,81		
	-0,72	15,26	0,02	15,26	-0,12	6,58	0,25	6,53		
	-0,37	14,50	0,57	14,76	0,36	7,56	1,05	7,81		
F(t) = 0,0625	-0,14	16,16	0,75	16,36	0,56	13,04	1,22	13,08	p = 0,3 et p = 0,6	
	0,25	21,34	0,78	21,09	-0,35	15,38	0,64	15,64		
	-1,39	11,45	-0,86	11,37	-0,35	15,38	0,64	15,64		
	-0,41	19,89	0,14	19,73	1,23	13,79	1,71	13,62		
	-0,10	20,25	0,37	19,89	0,29	8,97	0,78	8,88		
F(t) = 0,2500	-1,40	16,70	-0,49	16,89	-0,75	9,24	0,07	9,49	p = 0,3 et p = 0,8	
	0,71	17,78	1,03	17,57	0,91	15,07	1,34	15,04		
	F(t) = 0,9375									
	F(t) = 0,7500									
	F(t) = 0,5000									
Y	0,01	15,22	0,93	15,51	-1,24	8,64	-0,35	9,07	p = 0,5 et p = 0,6	
	-1,09	7,54	-0,56	7,69	-1,24	8,64	-0,35	9,07		
	-0,44	15,22	-0,08	14,99	-0,23	8,18	0,29	8,23		
	0,05	14,92	0,71	14,84	0,43	6,21	0,86	6,20		
	0,13	12,54	0,86	12,70	0,81	6,85	1,26	6,99		
F(t) = 0,0625	0,01	15,22	0,93	15,51	-1,24	8,64	-0,35	9,07	p = 0,5 et p = 0,8	
	-1,09	7,54	-0,56	7,69	-1,24	8,64	-0,35	9,07		
	-0,44	15,22	-0,08	14,99	-0,23	8,18	0,29	8,23		
	0,05	14,92	0,71	14,84	0,43	6,21	0,86	6,20		
	0,13	12,54	0,86	12,70	0,81	6,85	1,26	6,99		
F(t) = 0,2500	0,01	15,22	0,93	15,51	-1,24	8,64	-0,35	9,07	p = 0,5 et p = 0,8	
	-1,09	7,54	-0,56	7,69	-1,24	8,64	-0,35	9,07		
	-0,44	15,22	-0,08	14,99	-0,23	8,18	0,29	8,23		
	0,05	14,92	0,71	14,84	0,43	6,21	0,86	6,20		
	0,13	12,54	0,86	12,70	0,81	6,85	1,26	6,99		
F(t) = 0,7500	0,01	15,22	0,93	15,51	-1,24	8,64	-0,35	9,07	p = 0,5 et p = 0,6	
	-1,09	7,54	-0,56	7,69	-1,24	8,64	-0,35	9,07		
	-0,44	15,22	-0,08	14,99	-0,23	8,18	0,29	8,23		
	0,05	14,92	0,71	14,84	0,43	6,21	0,86	6,20		
	0,13	12,54	0,86	12,70	0,81	6,85	1,26	6,99		
F(t) = 0,9375	0,01	15,22	0,93	15,51	-1,24	8,64	-0,35	9,07	p = 0,5 et p = 0,8	
	-1,09	7,54	-0,56	7,69	-1,24	8,64	-0,35	9,07		
	-0,44	15,22	-0,08	14,99	-0,23	8,18	0,29	8,23		
	0,05	14,92	0,71	14,84	0,43	6,21	0,86	6,20		
	0,13	12,54	0,86	12,70	0,81	6,85	1,26	6,99		

et

$$RI = \left\{ \frac{1}{S} \sum_{s=1}^S [v_s(\hat{\theta}_j) - EQM(\hat{\theta}_j)]^2 \right\}^{1/2} / EQM(\hat{\theta}_j),$$

où le nombre de simulations exécutées est $S = 5\,000$ et l'EQM réelle $(\hat{\theta}_j)$ a été obtenue au moyen d'un ensemble indépendant de 50 000 simulations. Les estimateurs bootstrap de la variance se fondent chacun sur $B = 2\,000$ rééchantillons bootstrap. Nous obtenons les résultats de l'estimation de la variance de $\hat{\theta}_j$ représentant le total avec imputation et la fonction de distribution avec imputation en utilisant : i) l'estimateur bootstrap à demi-échantillon répété avec approximation de Monte Carlo appropriée, que nous représentons par v_{BRR} , comme dans l'équation (8) et avec approximation de Monte Carlo inappropriée obtenu en remplaçant $\hat{\theta}_{I(1)}^{(i)}$ par $\hat{\theta}_j^{(i)}$, que nous représentons par v_{B2} , et ii) l'estimateur par répliques équilibrées répétées (BRR) approprié, que nous représentons

par v_{BRR} , comme dans l'équation (9) et l'estimateur par répliques équilibrées répétées (BRR) inapproprié obtenu en remplaçant $\hat{\theta}_{I(1)}^{(i)}$ par $\hat{\theta}_j^{(i)}$, que nous représentons par v_{BRR2} . Le tableau 3 résume les résultats pour le biais relatif en pourcentage pour l'imputation aléatoire et pour l'imputation aléatoire corrigée. Nous ne présentons pas les résultats pour l'estimation du total de population Y , dans le cas de l'imputation aléatoire corrigée, parce que cette dernière élimine la variance due à l'imputation et que l'on peut donc appliquer des méthodes plus simples d'estimation de la variance (Chen et coll. 2000). Il est évident, si l'on considère le pourcentage élevé de biais relatif obtenu pour v_{B2} et v_{BRR2} , qu'il ne faut remplacer $\hat{\theta}_{I(1)}^{(i)}$ par $\hat{\theta}_j^{(i)}$ ni pour le bootstrap ni pour la méthode des répliques équilibrées répétées. Il est également évident que le biais qui entache les estimateurs de la variance, v_B et v_{BRR} , obtenus par le bootstrap à demi-échantillon répété et par la méthode des répliques équilibrées répétées est négligeable si les méthodes sont appliquées convenablement.

Tableau 1
Un ensemble de rééchantillons équilibrés construits d'après une MMOE

h	h						
	1	2	3	4	5	6	7
1	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)
2	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)
3	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
4	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
5	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
6	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
7	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
8	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
9	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
10	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)
11	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)
12	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)
13	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)
14	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)
15	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
16	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
17	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
18	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
19	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
20	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
21	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
22	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)
23	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)
24	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)

où $y_{hi} \sim N(\mu_h, \sigma_h^2)$ est indépendante de $e_{hik} \sim N(0, [1 - p] \sigma_h^2 / p)$ et les valeurs des paramètres sont

celles données dans le tableau 2. Pour une valeur particulière de la corrélation à l'intérieur des grappes, ρ , nous avons donc produit une population finie unique, puis nous

l'avons fixée et en avons tiré des échantillons répétés. Chaque simulation consistait à sélectionner $n_h = 2$ grappes avec remise dans la strate h pour $h = 1, \dots, H$ et à

dénombrer la grappe complète. Chaque unité finale de la grappe obtenue a été catégorisée indépendamment comme

étant un répondant ou un non-répondant avec les probabilités p et $(1 - p)$, respectivement, c'est-à-dire une réponse

uniforme. Ensuite, nous avons imputé les données pour les non-répondants par imputation aléatoire ainsi que par

imputation aléatoire corrigée, puis nous avons calculé le total de population et la fonction de distribution pour diverses valeurs de $F(t)$. Nous avons considéré deux valeurs de p , 0,1 et 0,3, ainsi que deux valeurs de ρ , 0,6 et 0,8. Notons que la fraction d'échantillonnage de premier degré est assez faible (0,064), si bien que les échantillonnages avec remise et sans remise sont essentiellement équivalents.

Pour comparer la performance des divers estimateurs de la variable, nous avons calculé le biais relatif (BR) en

$$\%BR = \frac{100}{S} \sum_{s=1}^S v_s(\hat{\theta}_s) / EQM(\hat{\theta}_s)$$

pourcentage et l'instabilité relative (IR) de chacun, définis respectivement comme étant

Tableau 2
Paramètres de la population finie

h	N _h	μ _h	σ _h	h	N _h	μ _h	σ _h
1	13	200	20,0	17	31	150	15,0
2	16	175	17,5	18	31	140	14,0
3	20	150	15,0	19	31	130	13,0
4	25	190	19,0	20	34	120	12,0
5	25	165	16,5	21	34	110	11,0
6	25	190	19,0	22	34	100	10,0
7	25	180	18,0	23	34	150	15,0
8	28	170	17,0	24	37	125	12,5
9	28	160	16,0	25	37	100	10,0
10	28	180	18,0	26	37	150	15,0
11	31	170	17,0	27	37	125	12,5
12	31	160	16,0	28	39	100	10,0
13	31	150	15,0	29	39	75	7,5
14	31	180	18,0	30	42	75	7,5
15	31	170	17,0	31	42	75	7,5
16	31	160	16,0	32	42	75	7,5

v_{BRR} est convergent pour les fonctions lisses des totaux en ont fait de même pour les estimateurs non lisses. On obtiendrait un estimateur BRR naïf pour les problèmes où les données sont imputées aléatoirement en remplaçant dans (9) $\theta_{(b)}$ et $\theta_{(c)}$ par $\theta_{(b)}^{(i)}$ et $\theta_{(c)}^{(i)}$ = $B^{-1} \sum_{j=1}^B \theta_{(b)}^{(i,j)}$, où $\theta_{(b)}^{(i,j)}$ est l'estimateur calculé pour X_j^i en utilisant les poids BRR. Mais cette solution produit des estimateurs de la variance non convergents, parce qu'elle ne tient pas compte de l'effet des données manquantes et de l'imputation aléatoire.

Pour appliquer correctement la méthode des répliques équilibrées répétées (BRR) en cas d'imputation aléatoire en procédant à une réimputation, nous devons tenir compte de ce que n_h est petit. Rappelons que, pour l'estimateur bootstrap, le fait que les n_h soient petits posait des difficultés, parce que la taille du nouvel échantillon de strate, $n_h - 1$, était plus petite que celle de l'échantillon de strate original, n_h . Il en est de même pour l'estimation par répétition équilibrée. Nous proposons une méthode simple pour contourner ces difficultés, au lieu d'obtenir la $b^{\text{ième}}$ répétition équilibrée répétée de l'estimateur, $\theta_{(b)}$, à partir de la même formule que pour θ , mais avec le poids $w_{hik(b)}$ égal à $2w_{hik}$ ou à 0 selon que la $(h)^{\text{ième}}$ grappe est sélectionnée ou non dans le $b^{\text{ième}}$ demi-échantillon, nous utilisons les poids originaux, mais nous incluons la $(h)^{\text{ième}}$ grappe soit deux fois soit pas du tout selon que cette grappe est sélectionnée ou non dans le $b^{\text{ième}}$ demi-échantillon. Si nous considérons les répliques équilibrées répétées (BRR) de cette façon, i) dans (9), l'estimateur résultant de la variance v_{BRR} ne change pas et ii) la taille du nouvel échantillon est la même que celle de l'échantillon original. Cette méthode des répliques équilibrées répétées peut être considérée comme une forme de bootstrap équilibré, mais il convient de souligner que l'estimation par le bootstrap équilibré décrite par Nigam et Rao (1996) dans le cas où aucune donnée ne manque ne peut être appliquée ici, car, même si elle s'appuie sur l'utilisation d'un nouvel échantillon de taille $n_h = 2$ dans chaque strate, elle le fait de façon telle qu'un rééchantillonnage reste encore nécessaire et ne peut donc être utilisée en cas d'imputation aléatoire.

1. Former l'ensemble de demi-échantillons, c'est-à-dire une unité par strate, au moyen d'une matrice de Hadamard comme décrit plus haut.
2. Obtenir la $b^{\text{ième}}$ répétition équilibrée répétée en répétant une deuxième fois chaque unité comprise dans le demi-échantillon obtenu. Représentons ceci par $\{y_{hi}^b: i = 1, \dots, n_h = 2\}$.

$$y_{hik} = y_{hi} + \varepsilon_{hik},$$

3. Poser que a_{hij}^* est l'indicateur de réponse associé à y_{hij}^* , $s_m^* = \{(h, i, j): a_{hij}^* = 0\}$, et $s_r^* = \{(h, i, j): a_{hij}^* = 1\}$. Appliquer la même méthode d'imputation que celle utilisée pour produire X_j aux unités de s_m^* , en utilisant les « répondants » compris dans s_r^* . Représenter la $b^{\text{ième}}$ répétition équilibrée répétée de X_j par $X_j^{t(b)}$.
4. Obtenir l'analogue par répliques équilibrées répétées $\theta_{(b)}^{(i)}$ de θ d'après l'ensemble de données obtenues par répliques équilibrées répétées avec imputation $X_j^{t(b)}$.
5. Répéter les étapes 1 à 4 pour chaque ligne de la matrice $B \times H$ pour obtenir $\theta_{(b)}^{(i)}$ pour $b = 1, \dots, B$ et appliquer la formule des répliques équilibrées répétées type (9) pour obtenir les estimateurs BRR, de la variance pour θ_p avec $\theta_{(c)} = B^{-1} \sum_{b=1}^B \theta_{(b)}^{(i)}$ (pour les mêmes raisons que celles exposées à la section 4, nous ne devrions pas remplacer $\theta_{(c)}$ par θ_j).

Nous pouvons étendre cette idée aux cas où $n_h > 2$ en utilisant la même méthode avec des demi-échantillons obtenus au moyen de multimatrices orthogonales équilibrées (MMOE) (Sitter 1993). Par exemple, le tableau 1 donne un ensemble de $B = 24$ rééchantillons équilibrés pour $H = 7$ strates avec $n_h = 4$ UPE dans chaque strate. Cet ensemble est obtenu en utilisant la MMOE donnée dans le tableau 1 de Sitter (1993) et en répétant une deuxième fois chaque unité rééchantillonnée, comme à l'étape 2 de la procédure décrite ci-dessus produit aussi un estimateur approximativement non biaisé de la variance. Les MMOE sont assez faciles à construire si n_h est un nombre pair en se servant de plans à blocs équilibrés incomplets et de matrices de Hadamard, mais sont difficiles à construire si n_h est un nombre impair. Elles permettent aussi de traiter les cas où la valeur de n_h n'est pas la même pour les différentes strates, mais leur construction devient alors nettement plus ardue (voir Sitter 1993).

6. UNE SIMULATION

Pour étudier les propriétés des estimateurs par rééchantillonnage proposés de la variance, nous avons considéré une population finie de $H = 32$ strates avec N_h grappes dans la strate h et 10 unités finales dans chaque grappe. Nous avons produit la caractéristique étudiée y_{hik} comme suit :

i) Si nous sélectionnons rééchantillon aléatoire simple de taille $m_h = (n_h - 1)/2$ sans remise et que nous répétons le tirage afin de tirer deux fois chaque unité, nous obtenons $n_h - 1$ unités. Si nous sélectionnons une unité supplémentaire au hasard à partir des $n_h - 1$ unités déjà rééchantillonnées, $\text{Var}^*(Y^*) = \sum_h (n_h + 3)s_h^2/n_h^2$.

ii) Si nous sélectionnons rééchantillon aléatoire simple de taille $m_h + 1$ sans remise et que nous répétons le tirage de sorte que chaque unité soit sélectionnée deux fois, nous aboutissons à $n_h + 1$ unités. Si nous écartons l'une de ces unités au hasard, $\text{Var}^*(Y^*) = \sum_h (n_h - 1)s_h^2/n_h^2$.

Donc, si nous utilisons la méthode (i) avec une probabilité de 1/4 et la méthode (ii) avec une probabilité de 3/4 lors de chaque itération bootstrap, nous obtenons le résultat souhaité. Cette méthode bootstrap à demi-échantillon répète donne des estimations de la variance approximativement non biaisées sans rééchantillonnement et la taille de l'échantillon bootstrap est égale à celle de l'échantillon original. Donc, si nous utilisons cet estimateur bootstrap original. Sitter (1996), tel que décrit plus haut, les estimateurs bootstrap résultants sont asymptotiquement non biaisés et convergents pour tout n_h , dans les conditions de régularité énoncées dans Shao et Sitter (1996) et dans Shao et coll. (1998).

4. APPROXIMATION DE MONTE CARLO APPROPRIÉE POUR LA VARIANCE BOOTSTRAP

Si, dans (5), v_B n'a aucune forme explicite, on peut utiliser l'approximation de Monte Carlo

$$v_B(\theta_j^*) \approx \frac{1}{B} \sum_{b=1}^B (\theta_j^{(b)*} - \bar{\theta}_j^*)^2, \quad (8)$$

où $\bar{\theta}_j^* = B^{-1} \sum_{b=1}^B \theta_j^{(b)*}$, $\theta_j^{(b)*} = \theta(X_j^{(b)*})$, et $X_j^{(b)*}, b = 1, \dots, B$ sont des ensembles de données bootstrap indépendants réimputés. En pratique, il est courant, dans de nombreuses applications bootstrap, de remplacer dans (8) la moyenne des estimateurs bootstrap θ_j^* par l'estimateur original θ_j (voir Rao et Wu 1985, page 232). Ce dernier est plus facile à utiliser et est donc le plus courant. Si aucune donnée n'est imputée, cette démarche est habituellement correcte. Par contre, il est fautif d'utiliser l'analogue pour remplacer l'estimateur bootstrap réimputé, parce que θ_j^* est le résultat d'une réalisation unique de l'imputation aléatoire, tandis que $\theta_j^* \approx E^*(\theta_j^*) \approx E_j'(\theta_j^*)$, puisque nous calculons la moyenne sur des réimputations répétées, et les valeurs de $\theta_j^{(b)*}$ et $E_j'(\theta_j^*)$ ne sont pas proches en cas d'imputation aléatoire. Si $\theta_j = Y_j$, on obtient, par exemple, $E_j'(Y_j^*) = Y_j^*$ donné à la

section 2 et la différence $Y_j' - Y_j$ n'est pas un terme relativement négligeable en cas d'imputation aléatoire. Donc,

$$v_{B2} = \frac{1}{B} \sum_{b=1}^B (\theta_j^{(b)*} - \bar{\theta}_j^*)^2 = \frac{1}{B} \sum_{b=1}^B (\theta_j^{(b)*} - \bar{\theta}_j^*)^2 + (\bar{\theta}_j^* - \theta_j^*)^2$$

5. UN BRR RÉPÉTÉ

Commençons par décrire l'application la plus courante des répétitions équilibrées répétées ou BRR, c'est-à-dire

lorsqu'aucune donnée ne manque. Un ensemble de B demi-échantillons, ou répétitions, équilibrés est formé en supprimant une grappe de premier degré de l'échantillon de chaque strate, où cet ensemble est défini par une matrice $B \times H$ (δ^{bh}) avec $\delta^{bh} = +1$ ou -1 selon que la première ou la deuxième grappe de premier degré de la strate h se trouve dans le $b^{\text{ème}}$ demi-échantillon et $\sum_{b=1}^B \delta^{bh} \delta^{bh'} = 0$ pour tout $h \neq h'$; autrement dit, les colonnes de la matrice sont orthogonales. Nous pouvons construire un ensemble minimal de demi-échantillons équilibrés B à partir d'une matrice de Hadamard $B \times B$ en choisissant n'importe quelle colonne H , sauf la colonne ne contenant que des valeurs $+1$, où $H + 1 \leq B \leq H + 4$. Représentons par $\theta_j^{(b)}$ l'estimateur d'enquête calculé d'après le $b^{\text{ème}}$ demi-échantillon. Nous pouvons obtenir l'estimateur $\theta_j^{(b)}$ en nous servant de la même formule que pour θ avec w_{hik} remplacé par $w_{hik}^{(b)}$ qui est égal à $2w_{hik}$ ou à 0 selon que la $(h_i)^{\text{ème}}$ grappe est sélectionnée ou non dans le $b^{\text{ème}}$ demi-échantillon. L'estimateur par répétitions équilibrées répétées, ou estimateur BRR, de la variance de θ est donné par

$$v_{\text{BRR}} = \frac{1}{B} \sum_{b=1}^B (\theta_j^{(b)} - \bar{\theta}_j^*)^2, \quad (9)$$

où $\bar{\theta}_j^* = \sum_{b=1}^B \theta_j^{(b)}/B$, et est souvent remplacée par θ . Krewski et Rao (1981) ont montré que l'estimateur de la variance

$$\begin{aligned} \text{Var}(\hat{Y}^I) &= \text{Var}\left[E_I(\hat{Y}^I)\right] + E\left[\text{Var}_I(\hat{Y}^I)\right] \\ &= \text{Var}\left(\frac{\sum_{s_r} \sum_{s_m} w_{hik} Y_{hik}}{s_r \sum_{s_r} w_{hik}}\right) + E\left(\hat{\sigma}^2 \sum_{s_m} w_{hik}^2\right), \quad (6) \end{aligned}$$

où

$$\hat{\sigma}^2 = \sum_{s_r} w_{hik} (Y_{hik} - \bar{Y}^I)^2 / \sum_{s_r} w_{hik},$$

De même, d'après (4),

$$\begin{aligned} \text{Var}^*(\hat{Y}^I) &= \text{Var}^*\left(\frac{\sum_{s_r} w_{hik}^* Y_{hik}^* \sum_{s_m} w_{hik}^*}{s_r \sum_{s_r} w_{hik}^*}\right) \\ &= E^*\left(\hat{\sigma}^{*2} \sum_{s_m} w_{hik}^{*2}\right) + E^*\left(\frac{\sum_{s_r} w_{hik}^* Y_{hik}^*}{\sum_{s_r} w_{hik}^*}\right)^2, \quad (7) \end{aligned}$$

où

$$\hat{\sigma}^{*2} = \sum_{s_r} w_{hik}^* (Y_{hik}^* - \bar{Y}^{I*})^2 / \sum_{s_r} w_{hik}^*,$$

En vertu de la théorie du bootstrap, les premiers termes du membre de droite de (6) et (7) convergent vers la même quantité, comme le font aussi $\hat{\sigma}^2$ et $\hat{\sigma}^{*2}$. Donc, l'estimateur bootstrap de Shao et Sitter est convergent si $\sum_{s_m} w_{hik}^{*2}$ et $\sum_{s_m} w_{hik}^2$ convergent vers la même quantité, ce qui est le cas si $n_h/n_h - 1$ converge vers 1 pour tous les h , car

$$E^*\left(\sum_{s_m} w_{hik}^{*2}\right) = E^*\left[\sum_{s_m} (1 - a_{hik}^*) w_{hik}^{*2}\right] = \sum_{s_m} (1 - a_{hik}) w_{hik}^2 / (n_h - 1).$$

Le deuxième terme du deuxième membre de (6) est la composante de la variance correspondant à l'imputation aléatoire, composante qui représente habituellement une faible part de la variance totale. Donc, la surestimation due à $n_h/(n_h - 1)$ n'est sérieuse que si les n_h sont très petits. Le cas où $n_h = 2$ est, néanmoins, un cas particulier important. Nous proposons maintenant une méthode bootstrap qui ne pose aucun problème dans le cas où la valeur des n_h est très faible, tout en demeurant valide de façon plus générale. Notons que nous utilisons un échantillon bootstrap de taille

$n_h - 1$ pour nous assurer que le premier terme du deuxième membre de (7) ait la même limite que le premier terme du deuxième membre de (6) (Rao et Wu 1988). En prenant n_h comme taille d'échantillon bootstrap dans la strate h , Rao et Wu (1988) ont montré que, dans le cas où il ne manque aucune donnée, l'estimateur bootstrap de la variance sous-estime cette dernière. Ils ont proposé un rééchantillonement pour contourner le problème, mais celui-ci ne produit pas d'estimateurs bootstrap corrects dans le cas de données imputées.

Dans le cas qui nous occupe, nous aurions idéalement besoin d'une méthode bootstrap où la taille de l'échantillon est égale à celle de l'échantillon original n_h qui produit un estimateur de la variance asymptotiquement non biaisé (dans le cas où il ne manque aucun donné) sans rééchantillonement. Nous allons montrer maintenant que l'on peut y arriver comme suit. Supposons qu'il ne manque aucune donnée et que toutes les valeurs de $n_h = 2m_h$ sont paires. Tirons un échantillon aléatoire simple de taille m_h sans remise, indépendamment, à partir de $\{Y_{hi}^* : i = 1, \dots, m_h\}$ et répétons chaque unité obtenue une seconde fois afin d'avoir $\{Y_{hi}^* : i = 1, \dots, n_h\}$. Nous donnons à cette méthode le nom de méthode bootstrap à demi-échantillon répété. L'estimateur résultant v_B sera alors approximativement non biaisé et convergent. Dans le cas linéaire où $Y^* = \sum_{h(k)} w_{hik} Y_{hik} = \sum_{h(k)} \sum_{i=1}^{n_h} Y_{hi}^* / n_h$ et $Y_{hi}^* = \sum_{k=1}^{m_h} n_h w_{hik} Y_{hik}$, la convergence de v_B découle de

$$\begin{aligned} \text{Var}^*(Y^*) &= \sum_{h(k)} \text{Var}^*(Y_{hi}^*) = \sum_{h(k)} \text{Var}^*\left(\frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}^*\right) \\ &= \sum_{h(k)} \text{Var}^*\left(\frac{2m_h}{n_h} \frac{1}{m_h} \sum_{i=1}^{m_h} Y_{hi}^*\right) \\ &= \sum_{h(k)} \text{Var}^*\left(\frac{1}{m_h} \sum_{i=1}^{m_h} Y_{hi}^*\right) \\ &= \sum_{h(k)} \text{Var}^*\left(\frac{1}{m_h} \sum_{i=1}^{m_h} Y_{hi}^*\right) \\ &= \sum_{h(k)} \text{Var}^*\left(\frac{1}{m_h} \sum_{i=1}^{m_h} Y_{hi}^*\right) \end{aligned}$$

c'est-à-dire l'estimateur approximativement non biaisé et convergent habituel de la variance où $s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (Y_{hi}^* - \bar{Y}^I)^2$. La convergence de v_B pour un θ_I non linéaire découle du cas linéaire et du développement en série de Taylor lorsque θ_I est une fonction des moyennes pondérées, ou des arguments présentés dans Shao et Rao (1994), Shao et Sitter (1996) et Shao et coll. (1998) lorsque θ_I est un estimateur non lisse, comme une médiane. Si la valeur de $n_h = 2m_h + 1$ est impaire, il est impossible de tirer un demi-échantillon exact. Le cas échéant, les deux résultats qui suivent nous mènent à une adaptation de l'idée qui précède.

3. BOOTSTRAP À DEMI-ÉCHANTILLON RÉPÉTÉ

En cas d'imputation pour remplacer des données manquantes, les estimateurs bootstrap naïfs de la variance obtenus en traitant l'ensemble de données ayant subi l'imputation, X_p , comme étant $X = \{y_{hik} : (h, i, j) \in s\}$, c'est-à-dire l'ensemble de données ne comportant pas de valeur manquante, ne reflète pas l'augmentation de la variance due à l'imputation et(ou) aux données manquantes, et produisent une sous-estimation importante. Par conséquent, ils ne sont pas convergents, car le fait de traiter simplement X_p comme étant X ne tient pas compte du processus d'imputation. Shao et Sitter (1996) l'ont fait remarquer et ont proposé de réaliser, sur l'ensemble de données bootstrap, une nouvelle imputation similaire à celle effectuée sur l'ensemble de données originales. La méthode bootstrap présentée dans Shao et Sitter (1996) peut être décrite comme suit.

1. Sélectionner un échantillon aléatoire simple $\{y_{hi}^* : i = 1, \dots, n_h - 1\}$ avec remise à partir de l'échantillon $\{y_{hi}^* : i = 1, \dots, n_h\}$, $h = 1, \dots, H$, indépendamment dans chacune des strates, où $y_{hi}^* = \{y_{hij}^* : (h, i, j) \in s\} \cup \{\tilde{y}_{hij}^* : (h, i, j) \in s_m\}$. Soit \tilde{a}_{hij}^* , l'indicateur de réponse associé à y_{hij}^* , $s_m^* = \{(h, i, j) : a_{hij}^* = 0\}$ et $s_r^* = \{(h, i, j) : a_{hij}^* = 1\}$. Appliquer la même méthode d'imputation que celle utilisée pour produire l'ensemble de données imputées X_p aux « non-répondants » compris dans s_m^* en utilisant les « répondants » compris dans s_r^* . Représenter l'analogue bootstrap de X_p par X_p^* .
2. Obtenir l'analogue bootstrap θ_p^* de θ , d'après l'ensemble de données bootstrap imputées X_p^* . Par exemple, si $\theta = Y$ dans (1) et $\theta_p = Y_p$ dans (2), alors

$$\theta_p^* = Y_p^* = \sum_{s_r^*} w_{hik}^* y_{hik}^* + \sum_{s_m^*} w_{hik}^* \tilde{y}_{hik}^* \quad (4)$$

$$v_B(\theta_p^*) = \text{Var}^*(\theta_p^*) \quad (5)$$

où Var^* représente la variance conditionnelle en rapport avec X_p^* , étant donné X_p .

Shao et Sitter (1996) montrent que l'estimateur bootstrap défini dans (5) est convergent pour les estimateurs θ lisses ainsi que non lisses. Dans le cas d'une méthode d'imputation aléatoire, l'une des conditions implicites du développement de ces estimateurs est que $n_h/(n_h - 1)$ tende vers 1. Le cas particulier où $\theta = Y$ permet d'illustrer ce point. Partant de (2), nous avons

hoteck pondérée décrite dans Rao et Shao (1992), que nous appelons tout simplement imputation aléatoire, et la méthode hordeck pondérée corrigée proposée par Chen, Rao et Sitter (2000) que nous appelons simplement ici imputation aléatoire corrigée. Nos résultats peuvent être facilement généralisés à l'imputation aléatoire avec résidus en présence de données auxiliaires (par exemple, imputation par régression aléatoire). La généralisation à d'autres formes d'imputation aléatoire pourrait être possible, mais ne sera pas envisagée ici.

Dans le cas de l'imputation aléatoire, des donneurs \tilde{y}_{hik} sont sélectionnés au hasard à partir de $\{y_{hik} : (h, i, j) \in s_r\}$ avec remise et avec probabilité w_{hik}^*/\tilde{T} , où $\tilde{T} = \sum_{s_r} w_{hik}$. Dans ce cas, $E_p(\tilde{X}_p) = (\tilde{S}/\tilde{T})U = Y_p$, c'est-à-dire un estimateur par quotient qui est asymptotiquement non biaisé et convergent pour X , où $\tilde{S} = \sum_{s_r} w_{hik} y_{hik}$. Ici, E_p représente l'espérance dans les conditions d'imputation aléatoire. La variance de X_p est plus forte que celle de Y_p à cause de l'imputation aléatoire. Cependant, la répartition des valeurs de réponse dans l'ensemble de données subsistant l'imputation. Cependant, les valeurs imputées résultantes doivent être des réalisations réelles.

Représentons l'estimateur avec imputation de la fonction de distribution dans les conditions d'imputation aléatoire par

$$F_p^*(t) = \left[\sum_{s_r} w_{hik} I(y_{hik} \leq t) + \sum_{s_m} w_{hik} I(\tilde{y}_{hik} \leq t) \right] / \tilde{U} \quad (3)$$

Pour obtenir l'estimateur avec imputation de la fonction de distribution dans les conditions d'imputation aléatoire corrigée, représentée par $F_p^*(t)$, nous remplaçons simplement \tilde{y}_{hik} dans (3) par \tilde{y}_{hik}^* . Dans le cas de l'estimation de la fonction de distribution, contrairement à celui de l'estimation du total, l'imputation aléatoire corrigée n'élimine pas la variance due à l'imputation. Cependant, Chen et coll. (2000) montrent qu'elle réduit considérablement la variance d'imputation comparativement à la méthode d'imputation aléatoire non corrigée. Les estimateurs $F_p^*(t)$ et $F_p^*(t)$ sont tous deux asymptotiquement non biaisés et convergents.

Pour étudier l'estimation de la variance liée aux méthodes de rééchantillonnage, nous supposons que n/N est négligeable, où $n = \sum n_h$, $N = \sum N_h$ et N_h est le nombre de grappes de premier degré dans la population.

Dans le présent article, nous poursuivons les travaux entrepris par Shao et Sitter (1996). En premier lieu, nous montrons à la section 3 comment leur méthode peut être modifiée pour traiter des strates de très petite taille (par exemple, deux UPE par strate). En deuxième lieu, nous discutons, à la section 4, de l'approximation de Monte Carlo appropriée pour les estimateurs bootstrap, question qui doit être résolue plus minutieusement qu'à l'ordinaire en cas de réimputation aléatoire. Celle-ci n'a aucun effet négatif sur les intervalles de confiance bootstrap fondés sur la méthode des centiles, mais, appliquée incorrectement, elle produit de mauvais résultats pour le t bootstrap. En troisième lieu, nous considérons une méthode d'estimation de la variance par répliques équilibrées répétées (BRR) avec une étape de réimputation qui peut être considérée comme une approximation analytique et symétrique de la méthode bootstrap. Enfin, nous présentons certains résultats d'étude en simulation pour étudier les propriétés de divers estimateurs bootstrap et BRR de la variance.

2. ÉCHANTILLONNAGE STRATIFIÉ À PLUSIEURS DEGRÉS ET IMPUTATION ALÉATOIRE

Bien que les méthodes exposées ici puissent être appliquées de façon plus générale, nous nous limitons au plan de sondage stratifié à plusieurs degrés utilisé couramment. Supposons que la population contienne H strates et que l'on sélectionne dans la strate h , u_h grappes indépendamment dans les diverses strates. Dans le cas d'une réponse complète à une question y , posons que

$$Y^h = \sum_{u_h=1}^{u_h} Y_{hi}^h / (n^h p_{hi}^h)$$

est un estimateur linéaire non biaisé du total de strate Y^h , où Y_{hi}^h est un estimateur linéaire non biaisé du total dans la grappe Y_{hi}^h pour une grappe sélectionnée d'après l'échantillonnage au deuxième degré et aux degrés subséquents. Représentons l'estimateur linéaire non biaisé du total, $Y = \sum Y^h$, par $Y = \sum Y^h$, qui peut s'écrire sous la forme

$$Y = \sum_{(hik)} w_{hik} y_{hik}^h \quad (1)$$

où s est l'échantillon complet d'unités et w_{hik} et y_{hik}^h représentent, respectivement, le poids d'échantillonnage et la valeur de la réponse donnée par la (hik) – ième unité échantillonnée. Souvent, on peut exprimer un estimateur d'enquête, θ , sous la forme de fonction d'un vecteur des totaux estimés, comme dans (1). Si on le souhaite, on peut estimer la fonction de distribution de la population par $\hat{F}^n(t) = \sum_{\sum w_{hik} I(Y_{hik}^h \leq t) / U}$, où $I(\cdot)$ est la fonction indicatrice habituelle et $U = \sum w_{hik}$. Certains estimateurs non lisses qui présentent un intérêt sont le p ème quantile d'échantillon,

$F^{-1}(p)$, où F^{-1} est la fonction quantile de F , et la proportion de personnes à faible revenu dans l'échantillon $\hat{F}^{-1}([1/2, F^{-1}(1/2)])$. Supposons que l'on observe la valeur y_{hik}^h pour $(hik) \in s_r$ et $s_r \subset s$, que nous appelons un échantillon manquant pour $(hik) \in s_m$, que nous appelons un non-répondant, avec $s = s_r \cup s_m$. En cas de données manquantes, il est courant d'utiliser $\{y_{hik}^h : (hik) \in s_r\}$ pour obtenir des valeurs imputées \tilde{y}_{hik}^h pour $(hik) \in s_m$; puis de traiter ces valeurs imputées comme s'il s'agissait d'observations réelles et d'estimer Y au moyen de

$$Y_I = \sum_{s_r} w_{hik} y_{hik}^h + \sum_{s_m} w_{hik} \tilde{y}_{hik}^h \quad (2)$$

L'imputation aléatoire consiste à imputer des valeurs pour remplacer les valeurs manquantes au moyen d'un échantillon aléatoire sélectionné parmi les répondants ou, s'il existe des données auxiliaires, en utilisant un échantillon aléatoire de résidus. Si l'imputation est réalisée convenablement, l'estimateur Y_I dans (2) est asymptotiquement non biaisé et converge, bien qu'il ne soit pas aussi efficace que Y dans (1). Dans tout l'article, nous supposons que,

dans chaque cellule d'imputation, la probabilité de réponse est constante pour une variable donnée, les situations de réponse pour diverses unités sont indépendantes et l'imputation est réalisée dans chacune des cellules d'imputation indépendamment des autres cellules, ou que,

Nous supposons aussi qu'existent les mêmes conditions asymptotiques que dans Shao et coll. (1998). Donc, la convergence (ou absence asymptotique de biais) s'entend de la même manière dans les conditions de Shao et coll. (1998), à mesure que la taille de l'échantillon de premier degré $n = \sum n_h$ augmente et tend vers l'infini. Il existe de nombreuses méthodes d'imputation aléatoire. Nous n'en considérons que deux ici, à savoir la méthode

Bootstrap à demi-échantillon répété et répliques aléatoires répétées

HIROSHI SAIGO, JUN SHAO et RANDY R. SITTER¹

RÉSUMÉ

Nous discutons de l'application du bootstrap avec une étape de réimputation en vue de tenir compte de la variance due à l'imputation (Shao et Sitter 1996) dans le cas d'un échantillonnage stratifié à plusieurs degrés. Nous proposons une méthode bootstrap modifiée qui ne nécessite pas de rééchantillonnage si bien que la méthode de Shao et Sitter peut être appliquée au cas de l'imputation aléatoire lorsque la taille de l'échantillon de strate de premier degré est très petite. La méthode que nous proposons est une méthode unifiée, applicable quelle que soit la méthode d'imputation (aléatoire ou non aléatoire), la taille de la strate (petite ou grande), le genre d'estimateur (lisse ou non lisse) ou le genre de problème (estimation de la variance ou estimation de la distribution d'échantillonnage). En outre, nous discutons de l'approximation de Monte Carlo qu'il convient d'utiliser pour la variance bootstrap lorsque l'on conjugue la réimputation à des méthodes de rééchantillonnage. Dans ces conditions, on doit agir plus prudemment qu'à l'ordinaire. Nous obtenons des résultats comparables pour la méthode des répliques équilibrées répétées qui est souvent utilisée dans le contexte des enquêtes et peut être considérée comme une approximation analytique du bootstrap. Enfin, nous présentons certains résultats d'étude en simulation afin d'examiner les propriétés de l'échantillon de taille finie et divers estimateurs de la variance applicables en cas d'imputation des données.

MOTS CLÉS : Hot deck; méthode des centiles; Monte Carlo; imputation; taille de l'échantillon bootstrap.

1. INTRODUCTION

Dans les enquêtes, la non-réponse à une question est un problème fréquent que l'on contourne habituellement en imputant une valeur pour compenser les données manquantes. Les diverses méthodes d'imputation appliquées en pratique se répartissent entre deux grands groupes : les méthodes d'imputation déterministes, telles que l'imputation par la moyenne, par le quotient ou par régression, ordinairement appliquées en se servant des données fournies par les répondants et de certaines données auxiliaires observées sur toutes les unités échantillonnées, d'une part, et l'imputation aléatoire, d'autre part. Dans les deux cas, l'imputation a souvent lieu à l'intérieur de classes d'imputation créées en se basant sur des variables auxiliaires. Le présent article porte sur l'imputation aléatoire. Généralement, l'imputation aléatoire est réalisée de sorte que l'application des formules d'estimation habituelles à l'ensemble de données imputées produise des estimateurs asymptotiquement non biaisés et convergents (par exemple, moyennes, totaux, quantiles). Des renseignements plus détaillés sur l'imputation aléatoire sont présentés à la section 2. Il est également courant, en pratique, de traiter les valeurs imputées comme des valeurs réelles lorsque l'on estime la variance des estimateurs appliqués aux données d'enquête. Toutefois, cette façon de faire cause une sous-estimation grave de la variance si la proportion de données manquantes est appréciable et produit de mauvais intervalles de confiance.

Certains moyens de surmonter cette difficulté ont été suggérés. Dans le cas de l'imputation aléatoire, Rubin (1978) ainsi que Rubin et Schenker (1986) ont proposé, pour tenir compte de l'augmentation de la variance, d'appliquer une méthode d'imputation multiple que l'on peut justifier dans une perspective bayésienne (Rubin 1987). Certains auteurs ont proposé des méthodes jackknife rajustées d'estimation de la variance pour l'imputation tant aléatoire que déterministe (Rao et Shao 1992; Rao 1993; Rao et Sitter 1995; Sitter 1997) dans des conditions d'échantillonnage stratifié à plusieurs degrés. Cependant, il est bien connu que l'on ne peut appliquer le jackknife à des estimateurs non lisses, comme un quantile d'échantillon ou une proportion estimative de faible revenu (Mantel et Singh 1991). Deux méthodes sont applicables aux estimateurs aussi bien lisses que non lisses pour tenir compte des données imputées aléatoirement, à savoir la méthode corrigée des répliques équilibrées répétées, ou BKR pour *Balanced repeated replication*, proposée par Shao, Chen et Sitter (1996) (voir aussi Efron 1994) avec une étape de réimputation pour tenir compte de la variance due à l'imputation. La méthode bootstrap demandée plus de calculs, mais est facile à justifier et à comprendre, et représente une méthode unifiée qui est applicable quelle que soit la méthode d'imputation (aléatoire ou non aléatoire), le genre d'estimateur $\hat{\theta}$ (lisse ou non lisse) ou le genre de problème (estimation de la variance ou estimation de la distribution d'échantillonnage).

¹ Hiroshi Saigo, School of Political Science and Economics, Waseda University, 1-6-1 Nishitwaseba Shinjuku, Tokyo, 169-8050 Japan; Jun Shao, Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706, USA; Randy R. Sitter, Professor, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

l'estimation ponctuelle ainsi que l'estimation de la variance, souvent à peu de frais. Singh et coll. (1994) ont proposé d'augmenter le nombre d'UPB pour contrôler la taille des échantillons dans les régions non plantifiées. Puisque, comme l'a fait remarquer Fuller, le client en demandera toujours plus que ce qui est spécifié au stade de la conception de l'enquête, il est impossible de prévoir la liste complète des régions présentant un intérêt. Toutefois, si l'on augmente le nombre d'UPB, on augmente la probabilité de recueillir des données réelles pour des domaines analytiques non prévus.

Kalton (1994) a avancé une deuxième raison d'augmenter le nombre d'UPB par région. Selon lui, cette mesure rendrait les estimations de la variance nettement plus stables. Il en est ainsi même pour de très grandes enquêtes nationales comptant de nombreuses UPB. La NHS a été remaniée en 1995 en vue de faire passer le nombre d'UPB de 196 à 359. De ces 359 UPB, 264 étaient des UPB choisies sans certitude. Toutefois, sept étaient seulement comptaient plus de huit UPB de ce type. Bien que l'estimation directe de la variance au niveau de l'état demeure problématique pour la plupart des états, il y a maintenant plus de possibilités de calculer une estimation moyenne de la variance pour des groupes d'états ayant des caractéristiques communes, au lieu de devoir regrouper tous les états pour calculer une moyenne nationale.

7. SOMMAIRE

Les méthodes indirectes d'estimations régionales seront toujours nécessaires, puisque l'on ne connaît jamais d'avance l'ensemble complet des domaines d'analyse. La demande d'estimations régionales augmente partout dans le monde. Cependant, de nombreuses mesures peuvent être prises à l'étape de l'élaboration du plan de sondage pour améliorer les estimations régionales directes, qu'il s'agisse d'estimations ponctuelles ou d'estimations de la variance. Ces étapes incluent la stratification conformément aux domaines d'analyse connus, le suréchantillonnage au niveau régional et l'augmentation du nombre d'UPB. Selon

BIBLIOGRAPHIE

- le genre de données, il est souvent possible de regrouper des données recueillies pour plusieurs années de référence, d'utiliser des données recueillies dans le cadre d'autres enquêtes en fonction desquelles les questions ont été unifor-misées ou celles recueillies par des méthodes d'estimations fondées sur une base de sondage double. Ces mesures réduisent le besoin d'estimations indirectes et augmentent l'exactitude de ces estimations lorsqu'elles sont nécessaires.
- CHROMY, J.R., BOWMAN, K.R. et PENNE, M.A. (1999). The National Household Survey on Drug Abuse Sample Design Plan. Préparé pour Substance Abuse and Mental Health Services Administration, Rockville Maryland.
- CITRO, C.F. et KALTON, G. (2000). Small-area Estimates of School-age Children in Poverty: Evaluation of Current Methodology. National academy press, Washington, D.C.
- FULLER, W.A. (1999). Environmental surveys over time. *Journal of agricultural, Biological, and Environmental Statistics*, 4, 331-345.
- GHOSH, M. et RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.
- KALTON, G. (1994). Commentaires sur l'article de Singh, Gambino et Mantel. *Techniques d'enquête*, 20, 19-21.
- MARKER, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.
- RAO, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle. *Techniques d'enquête*, 25, 199-212.
- SINGH, M.P., GAMBINO, J. et MANTEL, H.J. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*, 20, 3-15.
- SIRKEN, M.G. et MARKER, D.A. (1993). Dual frame sample surveys based on NHIS and state RDD surveys. *Proceedings of the 1993 Public Health Conference on Records and Statistics*.

Tableau 4
Nombre d'états pour lesquels on peut atteindre un c.v. de 30 %, 20 % ou 10 % au moyen de l'échantillon régional de la NHS de 1995 uniquement, de l'estimateur à base de sondage double non biaisé, ou d'un supplément CA, ou pour lesquels on ne peut atteindre le c.v. visé, pour quatre variables particulières

c.v. Sources des données	Proportion non				Proportion de
	assurée :	assurée :	moins de 19 ans	faible revenu, enfants	18 ans et plus
30 % Echant. régional uniquement	42	31	31	31	46
Avec supplément CA	9	20	19	19	5
Impossible de satisfaire l'exigence	0	0	1	1	0
10 % Echant. régional uniquement	32	15	10	10	37
Avec supplément CA	19	35	40	40	14
Impossible de satisfaire l'exigence	0	1	1	1	0
10 % Echant. régional uniquement	8	2	2	2	15
Avec supplément CA	40	41	39	39	36
Impossible de satisfaire l'exigence	3	8	8	10	0

Tableau 5

Nombre d'états pour lesquels il est possible d'atteindre une VRRCEQM de 10 % au moyen de l'échantillon régional de la NHS de 1995 uniquement ou d'un supplément CA ou pour lesquels il est impossible d'atteindre la valeur visée, pour quatre variables particulières

Source des données	Proportion non				Proportion de
	assurée :	assurée :	moins de 19 ans	faible revenu, enfants	18 ans et plus
Echant. régional uniquement	8	2	2	2	15
Avec supplément CA					
Estimateur non biaisé	40	41	39	39	36
Impossible de satisfaire l'exigence	30	47	49	49	35
Estimateur non biaisé	3	8	10	10	0
Estimateur biaisé	13	2	0	0	1

Pour les caractéristiques pour lesquelles les attentes des ménages abonnées et non abonnées au téléphone diffèrent, l'utilisation d'un estimateur non biaisé fondé sur une base de sondage double produit des estimations plus exactes pour les régions où le taux de pénétration du téléphone est élevé. Pour celles où ce taux est faible, les estimations produites pour les caractéristiques pour lesquelles les attentes des ménages abonnées et non abonnées au téléphone sont les mêmes sont plus exactes si l'on utilise un estimateur à base de sondage double éventuellement biaisé. Le choix de l'estimateur à base de sondage double approprié pour une région et une caractéristique données permet de produire des estimations exactes pour une forte proportion de régions.

6. AMÉLIORATION DES ESTIMATIONS PONCTUELLES ET DE LA VARIANCE

très importante lorsque la stratification ne concorde pas avec les domaines d'analyse. L'utilisation de totaux de pas sous-populations (par exemple, variables démographiques) au niveau des régions. Cependant, il n'est pas possible de contrôler un aussi grand nombre de sous-populations au niveau régional qu'au niveau national, étant donné la taille plus petite des échantillons.

Il existe aussi de nombreuses méthodes permettant d'améliorer les estimations de la variance au niveau régional. Habituellement, le nombre d'UPB sélectionnées dans une région donnée est très faible. Par conséquent, le nombre de degrés de liberté pour l'estimation de la variance entre UPB (ou de la variance totale) est faible. Une solution consiste à calculer la moyenne des estimations de la variance pour les diverses régions, mais elle masque le fait que les estimations sont, en général, de qualité nettement meilleure pour certaines régions que pour d'autres. Une autre solution consiste à utiliser des fonctions généralisées de la variance pour lisser les estimations de cette dernière. Une meilleure méthode consiste à régler le problème de la variance des estimations régionales lors de l'établissement du plan de sondage. L'augmentation du nombre d'UPB, conjuguée à une réduction de la taille de l'échantillon dans chaque UPB, améliore significativement

4. UNIFORMISATION DES ENQUÊTES

Un autre moyen peu coûteux d'améliorer la qualité des estimations consiste à uniformiser les questions d'une enquête à l'autre. Eurostat a fait beaucoup d'efforts en vue d'harmoniser plusieurs enquêtes à l'échelon tant national qu'international. L'European Community Household Panel Survey (ECHP) vise à recueillir des renseignements cohérents sur les divers pays membres. L'organisme a également entrepris l'uniformisation de l'Enquête sur la population active de chaque pays membre. Cette uniformisation rend les données plus comparables entre pays.

L'uniformisation des enquêtes réalisées auprès d'une même population augmente la taille des échantillons, donc permet de produire des estimations régionales de meilleure qualité. Statistiques Finland est en train d'uniformiser le processus de collecte de données sur le revenu et d'autres variables pour ses diverses enquêtes. La Permanent Survey on Living Conditions de Statistiek Netherlands s'appuie sur une procédure commune pour recueillir les renseignements de base dans le cadre d'une série d'enquêtes sociales.

Même si l'énoncé des questions est le même pour les diverses enquêtes, les données ne sont pas toujours entièrement comparables. Le mode de collecte des données choisis, ainsi que l'ordre des questions peuvent causer des différences (Groves 1989).

5. ESTIMATION D'APRÈS UNE BASE DE SONDAGE DOUBLE

Parfois, il est possible de compléter les données d'une enquête réalisée sur place par des données recueillies par téléphone, donc d'augmenter la taille des échantillons régionaux à plus faible coût. L'enquête hollandaise sur la demande de logements est un exemple national avec interview sur place. Pour produire les estimations régionales, un supplément téléphonique est réalisé pour plus de 100 municipalités. Le tableau 3 montre la taille des

Municipalité	Enquête nationale sur place	Supplément téléphonique	Total
Leek	56	569	625
Marum	29	299	328
Stochteren	44	456	500
Zuidhorn	54	558	612
Emmen	770	224	994
Avereest	134	465	599
Bathmen	24	506	530
Dalfsen	157	466	623
Deventer	316	335	651
Diepenveen	47	336	383

Tableau 3
Nombre de répondants, base de sondage double, pour certaines municipalités, enquête hollandaise sur la demande de logements

échantillons pour l'enquête nationale sur place, le supplément téléphonique et l'échantillon total pour 10 municipalités.

Sirken et Marker (1993) ont décrit l'estimation fondée sur une base de sondage double pour la U.S. National Health Insurance Survey (NHIS) en se basant sur le plan de sondage en vigueur de 1985 à 1994. Le tableau 4 présente les résultats correspondants pour le plan de sondage courant, adopté en 1995. On y compare la capacité de produire des estimations au niveau de l'état d'après les données de l'enquête nationale recueillies sur place et d'après des estimations non biaisées fondées sur une base de sondage double comptant un nombre illimité d'interviews téléphoniques supplémentaires. Il faut jusqu'à 100, 200 et 2 000 interviews téléphoniques par état pour atteindre un c.v. de 30 %, 20 % et 10 %, respectivement. Si la proportion de ménages non abonnés au téléphone est forte dans une région, il se pourrait qu'aucun nombre d'interviews téléphoniques supplémentaires ne soit suffisant pour produire des estimations non biaisées ayant l'exactitude souhaitée.

Le cas échéant, on ne peut atteindre le degré souhaité d'exactitude qu'en utilisant un estimateur éventuellement biaisé qui regroupe toutes les données, indépendamment de leur mode de collecte. Il faut alors calculer la valeur relative de la racine carrée de l'erreur quadratique moyenne (VRCEQM) au lieu du coefficient de variation pour évaluer l'exactitude des estimations. Cependant, pour certaines caractéristiques, les ménages abonnés au téléphone ont d'autres attentes que ceux qui n'ont pas le téléphone. Le biais résultant peut, de nouveau, empêcher d'obtenir le niveau souhaité d'exactitude. Pour estimer le biais qui entache chacune de ces variables, nous avons comparé les réponses à la NHIS des ménages ayant et n'ayant pas le téléphone. Le tableau 5 montre comment le nombre d'états pour lesquels il est possible d'atteindre une VRCEQM de 10 % varie selon la question, une fonction du biais pour les ménages abonnés au téléphone et le taux de pénétration dans chaque état.

à celle pour l'état le plus petit était de 1/1 pour la CPS, de 60/1 pour la SIPP et de 110/1 pour la NHIS. Les ratios correspondants pour les coefficients de variation étaient de 3,5/1, 7,5/1 et 10,5/1, respectivement. Le surechantillonnage a réduit presque des deux tiers les coefficients de variation pour les états les plus petits.

Il ne faut pas perdre de vue que le surechantillonnage fondé sur les caractéristiques géographiques ne réduit pas nécessairement la variabilité dans d'autres domaines étudiés, comme les sous-groupes démographiques. Dans le cas de la CPS, le ratio de la taille de l'échantillon du plus grand état à celle du plus petit était de 15/1 pour les enfants, de 20/1 pour les personnes âgées, de 500/1 pour les Noirs et de 800/1 pour les Hispaniques.

Dans le cas de la U.S. National Employer Health Insurance Survey (NEHS) de 1994, on a essayé d'équilibrer le surechantillonnage au niveau des états de façon à pouvoir produire des estimations exactes au niveau de l'état et au niveau national. L'échantillon global de 40 000 échantillonnements a dû être réparti entre les 51 états afin de pouvoir produire des estimations directes par état. Trois options ont été envisagées.

- Option A : La répartition optimale nationale (fondée sur l'emploi total dans l'état) a produit un échantillon de très petite taille dans certains états.
- Option B : La répartition égale entre tous les états a produit des estimations nationales inefficaces.
- Option C : Au moins 400 questionnaires remplis par état (répartition d'après le nombre d'employés à la puissance 0,3).

Le ratio correspondant du c.v. pour l'état le plus grand au c.v. pour l'état le plus petit était de 7,2/1 pour l'option A, 1/1 pour l'option B et 1,8/1 pour l'option C. Comparative-

ment à l'option C, le c.v. des estimations nationales était

Proportion non assurée :		Proportion non assurée :		Proportion non assurée :		Proportion de fumées : 18 ans et plus	
1 année	42	31	28	45	42	45	45
2 années	46	35	36	50	46	50	50
3 années	49	41	37	51	49	51	51
c.v. de 20 %	31	13	10	36	31	36	36
1 année	36	29	24	44	36	44	44
2 années	42	31	31	46	42	46	46
c.v. de 10 %	7	2	2	14	7	14	14
1 année	14	3	3	25	14	25	25
2 années	22	7	4	32	22	32	32
3 années							

Tableau 2
Sommaire du nombre d'états (sur 51) pour lesquels la taille de l'échantillon de la NHIS de 1995 est suffisante pour atteindre un c.v. de 30 %, 20 % ou 10 % par agrégation des données de plusieurs années de référence pour quatre variables

effectives de l'échantillon. L'un des inconvénients de la combinaison de données de plusieurs années tient au fait que les estimations produites ne permettent pas de déceler rapidement les variations au cours du temps. Donc, si l'objectif principal est d'obtenir des données chronologiques, il faut appliquer d'autres méthodes pour augmenter la taille de l'échantillon. Le tableau 2 montre, pour la NHIS de 1995, le nombre d'états pour lesquels il est possible d'atteindre divers niveaux d'exactitude par regroupement des données de deux ou trois années d'enquête. L'agrégation permet manifestement d'obtenir des coefficients de variation de 30 % et de 20 %. Toutefois, pour nombre d'états, même l'agrégation de trois années de données ne permet pas d'atteindre un coefficient de variation de 10 %.

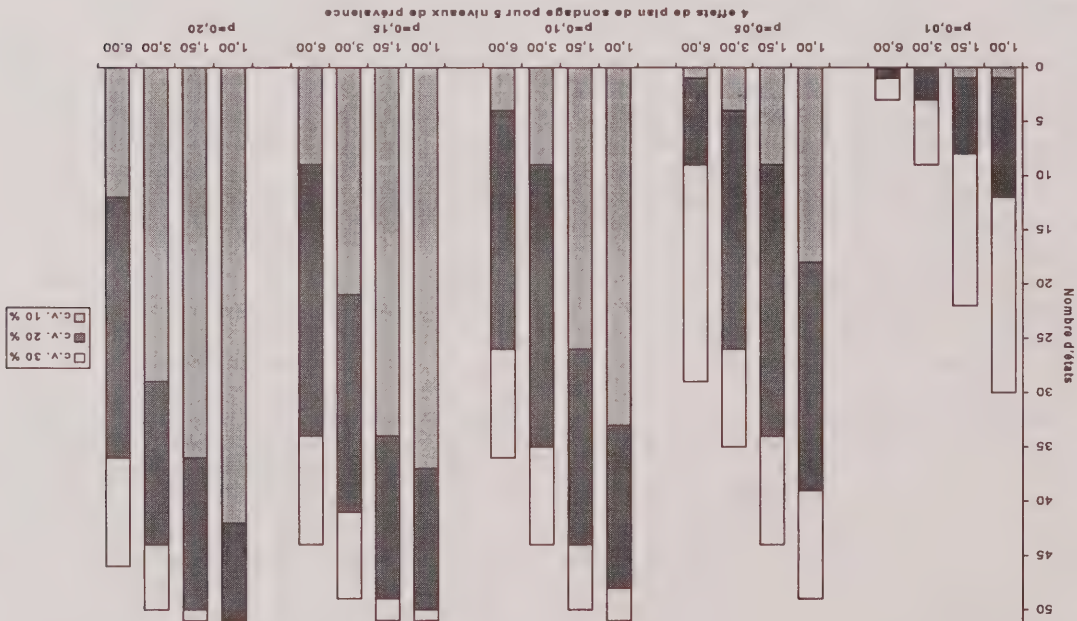
3. COMBINAISON DES DONNÉES DE PLUSIEURS ANNÉES

Un moyen peu coûteux d'augmenter la taille des échantillons régionaux consiste à regrouper les données de plusieurs cycles d'une enquête répétée. Combiner les données de k années d'une enquête annuelle n'augmente pas vraiment la taille effective de l'échantillon d'un facteur k parce qu'ordinairement, les cycles annuels consécutifs d'une enquête sont réalisés dans les mêmes unités primaires d'échantillonnage (UPB), voire même pour des segments de région adjacents. Il existe donc une certaine corrélation entre les données annuelles qui réduit sensiblement la taille effective de l'échantillon.

national.

17 % plus faible dans le cas de l'option A, mais 22 % plus élevé dans le cas de l'option B. L'option C a été choisie de préférence à l'option A puisqu'elle réduit les coefficients de variation au niveau de l'état d'un facteur 4 en n'augmentant que modérément le coefficient de variation au niveau

Figure 1. Nombre d'états répondant au critère de c.v. pour la NHIS de 1995 (44 000 ménages, 100 000 personnes)



Le suréchantillonnage permet d'augmenter de façon significative l'exactitude des estimations régionales directes en ne réduisant que très faiblement celle des estimations nationales. À titre d'exemple simple, considérons une enquête nationale réalisée auprès de 5 000 personnes mais pour laquelle, compte tenu du plan d'échantillonnage aléatoire, 10 des régions étudiées ne contiennent chacune que 100 cas. Une solution consisterait à doubler l'effectif de l'échantillon pour le porter à 200 pour chacune de ces régions, tout en maintenant à 5 000 la taille de l'échantillon national. La taille effective de l'échantillon utilisé pour calculer les estimations nationales serait réduite à cause de ce suréchantillonnage, mais resterait supérieure à 4 000, si bien que le coefficient de variation des estimations nationales augmenterait de moins de 10 %. En revanche, le coefficient de variation des estimations produites pour chacune des 10 régions diminuerait de 30 % parce que l'effectif de l'échantillon a doublé.

Depuis 1999, pour la U.S. National Household Survey on Drug Abuse, on recourt à une combinaison de stratification et de suréchantillonnage pour produire des estimations directes par état (Chromy, Bowman et Penne 1999).

Singh et coll. (1994) ont donné un exemple de suréchantillonnage régional dans le cas de l'Enquête sur la population active du Canada. En tout, 70 % de l'échantillon ont été répartis de façon à produire des estimations

nationales et provinciales optimales. Les 30 % restants ont été utilisés pour compléter les échantillons régionaux en vue d'améliorer l'exactitude des estimations à ce niveau. Ce compromis à l'étape du plan de sondage a fait augmenter de 10 % à 20 % le coefficient de variation des estimations nationales du taux de chômage, mais a réduit d'une valeur allant jusqu'à 50 % celui des estimations régionales.

Un plan de sondage comparable a été utilisé au Danemark pour l'Enquête sur la santé et la morbidité de 2000. L'enquête a été réalisée auprès de deux échantillons nationaux comptant chacun 6 000 personnes. Un échantillon supplémentaire de 8 000 personnes a été réparti de façon à obtenir au moins 1 000 répondants par comté.

Pour illustrer l'effet du suréchantillonnage sur les coefficients de variation, on peut aussi comparer la CPS de 1996 et la NHIS de 1995 à la Survey of Income and Program Participation (SIPP) réalisée en 1996 aux États-Unis. Le plan d'échantillonnage de la CPS prévoit non seulement une stratification selon l'état, mais aussi un suréchantillonnage des petits états. Celui de la NHIS prévoyait la stratification selon l'état, mais non le suréchantillonnage selon les caractéristiques démographiques (les groupes de minorités visibles ont été suréchantillonnés, mais ils ont tendance à être établis dans les états les plus peuplés). En revanche, le plan de sondage de la SIPP ne prévoyait ni stratification selon l'état ni suréchantillonnage. Le ratio de la taille de l'échantillon pour l'état le plus grand

L'optimisation de la stratification et du surechantillonnage ainsi que régionales devrait aussi être vue comme un compromis. Accepter de réduire dans une certaine mesure l'exactitude des estimations nationales permet souvent d'améliorer considérablement celle de nombreuses estimations régionales. Dans certains cas, on peut alors élaborer un plan de sondage permettant de produire ces estimations régionales exactes par une méthode directe. Pour d'autres, il faudra continuer d'utiliser des modèles, mais la stratification pourrait permettre de produire des estimations non biaisées (mais variables) intégrables dans les estimations basées sur le modèle. Comme l'illustre l'exemple qui suit, isolément, la stratification est utile, mais n'améliore que de façon limitée les estimations régionales.

La Current Population Survey (CPS) réalisée aux États-Unis par le Census Bureau comporte une stratification selon l'état et le taux de chômage depuis 1985. Par contre, l'échantillon de la United States National Health Interview Survey (NHIS), une autre grande enquête du Census Bureau, a été stratifié selon la région, la situation de région métropolitaine, les données sur la population active, le revenu et la composition raciale jusqu'en 1994. Les tailles résultantes d'échantillons pour les divers états variaient d'année en année et ne permettaient pas de produire des estimations non biaisées au niveau de l'état. En raison de l'échantillonnage aléatoire, de 1985 à 1994, deux états n'ont pas été représentés dans l'échantillon de la NHIS, ce qui ne se serait pas produit en cas de stratification selon l'état.

À partir de 1995, la stratification de l'échantillon de la NHIS a été réalisée selon l'état et la situation de région métropolitaine. Le tableau 1 donne le nombre d'états pour lesquels la taille de l'échantillon lors de la NHIS de 1995 était suffisante pour produire des estimations exactes à divers niveaux de détail pour quatre mesures importantes de la santé. Les interviews de la NHIS sont réalisées auprès d'environ 44 000 ménages comptant, en tout, environ

Tableau 1
Résumé du nombre d'états (sur 51, y compris le district fédéral de Columbia) pour lesquels la taille de l'échantillon de la NHIS de 1995 est suffisante pour atteindre un c.v. de 30 %, 20 % ou 10 % pour quatre variables particulières (44 000 ménages, 100 000 personnes)

Coefficient de variation (c.v.)	Proportion non assurée : tous âges confondus (p = 13,5 %)	Proportion non assurée : moins de 19 ans (p = 12,2 %)	Proportion non assurée : faible revenu, enfants (p = 20,4 %)	Proportion de fumeurs : 18 ans et plus (p = 25,2 %)
30 %	42	31	28	45
20 %	31	13	10	36
10 %	7	2	2	14

Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects

DAVID A. MARKER¹

RÉSUMÉ

Habituellement, les enquêtes nationales sont conçues pour produire des estimations pour le pays dans son ensemble et pour les principaux niveaux géographiques. Cependant, la demande d'estimations régionales pour les mêmes variables que celles évaluées dans le cadre de ces enquêtes ne cesse de croître. Par exemple, de nombreux pays en transition sont en train d'abandonner le processus centralisé de prise de décision et les pays occidentaux, comme les États-Unis, mettent sur pied des programmes, tels que celui du bien-être, dont la responsabilité est transférée de l'administration fédérale aux états. Souvent, les estimations régionales calculées d'après les données d'enquêtes nationales sont trop instables pour être utiles, ce qui pousse à rechercher des moyens de les améliorer. Bien que l'on puisse toujours produire des estimations régionales indirectes, dépendantes d'un modèle, il est préférable de produire des estimateurs directs dans la mesure du possible. La stratification et le suréchantillonnage permettent d'augmenter le nombre de régions pour lesquelles il est possible de calculer des estimations directes exactes. Pour d'autres régions, on peut recourir à certaines formes d'estimations fondées sur une base de sondage double de façon à combiner les données de l'enquête nationale à celles de suppléments d'enquête réalisés dans des régions particulières en vue de produire des estimations directes. Dans le présent article, on passe en revue les méthodes qui peuvent être utilisées pour produire des estimations régionales directes.

MOTS CLÉS : Estimations régionales; estimations directes; stratification; suréchantillonnage; estimations d'après une base de sondage double.

1. INTRODUCTION

Partout dans le monde, la demande d'estimations régionales augmente. Au cours des années 1990, les pays en transition ont abandonné le processus centralisé de prise de décision et, pour ce faire, ont eu besoin d'estimateurs exactes des conditions économiques et démographiques locales. Aux États-Unis, l'administration fédérale a transféré la responsabilité de nombreux programmes sociaux aux 50 états. Or, pour évaluer le succès de ces efforts, des estimations exactes sont nécessaires pour chaque état. Ainsi, les données de certains programmes, comme celui des estimations régionales du revenu et de la pauvreté (Citro et Kalton 2000), doivent être produites à un niveau de détail géographique nettement plus fin, par exemple, pour des milliers de districts scolaires. Aussi bons que soient les plans de sondage établis par les concepteurs d'enquête, comme le soutient Fuller, le client aura toujours besoin de plus de renseignements qu'il ne l'est spécifié au stade de l'élaboration du plan de sondage (Fuller 1999, page 344). Idéalement, ces estimations régionales devraient être produites au moyen d'estimateurs directs (axés sur le plan de sondage). Malheureusement, aux faibles niveaux d'agrégation, les estimations directes sont trop instables pour être publiées et/ou utilisées comme fondement de l'élaboration de politiques. Par conséquent, de nombreux chercheurs se sont efforcés de mettre au point diverses méthodes d'estimation indirectes (Marker 1999; Rao 1999; Ghosh et Rao 1994).

Pour toute enquête nationale, décider du plan optimal de stratification et de suréchantillonnage se résume à faire un compromis entre nombre de variables étudiées.

2. STRATIFICATION ET SURÉCHANTILLONNAGE

Ici, nous abordons le problème sous un angle différent et cherchons à réduire au minimum le recours à des modèles jamais possible d'anticiper toutes les utilisations des données d'une enquête, ni d'attribuer une taille d'échantillon suffisante à tous les domaines observés, si bien qu'il faudra toujours utiliser des estimateurs indirects. Toutefois, l'est possible de faire, au stade de la conception de l'enquête, des choix qui augmentent considérablement la capacité de produire des estimations régionales directes d'après les données d'enquêtes nationales, choix qui pourraient aussi augmenter la capacité d'utiliser les données de ces enquêtes pour produire des estimations indirectes, au besoin. Le présent article est une mise à jour de l'excellent document traitant du même sujet publié par Singh, Gambino et Mantel (1994). Les questions ayant trait au plan de sondage examinées ici incluent la stratification et le suréchantillonnage, la combinaison de données recueillies pour plusieurs années de référence, l'uniformisation des enquêtes, l'estimation fondée sur une base de sondage double et l'évaluation de l'exactitude des estimations.

¹ David A. Marker, Westat 1650 Research Blvd., Maryland, U.S.A. 20850. Courriel électronique: DavidMarker@Westat.com.

nouveaux arrivants dans la population. Notons qu'en l'absence d'un échantillon de remise à niveau, les nouveaux arrivants ne sont représentés dans les panels que par les cohabitants absents au départ. Il se pourrait que le calage des poids appliqués à l'échantillon combiné sur les totaux de population pour chacun des domaines temporels (lorsqu'on peut préciser quelles sont les unités du panel provenant de ces domaines) ne soit ni faisable ni judicieux pour les raisons déjà mentionnées à la section 3.1.

6. RÉSUMÉ ET CONCLUSIONS

Les méthodes de pondération décrites dans le présent article peuvent être utilisées pour combiner les renseignements provenant de plusieurs panels d'une enquête-ménage répétée afin de produire des estimations transversales dans des conditions assez générales ayant trait à des panels dont le plan de sondage est donné; les questions ayant trait à la détermination des fractions optimales d'échantillonnage pour les panels, parallèlement à la combinaison efficace des données de panels dépassent le cadre de l'article. Nous

avons montré que, si une enquête à panels multiples peut être considérée comme un cas spécial d'enquête à bases de sondage multiples, ses caractéristiques dynamiques distinctives rendent problématique, voire impossible, l'application des méthodes classiques utilisées pour les bases de sondage multiples. Les méthodes de pondération proposées, qui tiennent compte de la dynamique de la population et des panels, comprennent un rajustement simple des poids de chaque panel qui est proportionnel à la taille effective des panels respectifs. Ces méthodes sont commodées du point de vue opérationnel, quel que soit le nombre de panels chevauchants et pour diverses possibilités de définition des domaines temporels de panel. Nous avons également étudié les questions théoriques et pratiques que posent le rajustement par calage et l'intégration des diverses méthodes de pondération utilisées dans le cas d'une enquête à panels multiples. Plus précisément, nous soutenons que le rajustement des poids en vue de combiner les panels devrait précéder le rajustement par la méthode du partage des poids et que le rajustement final des poids devrait être celui fait par calage. Une étude empirique détaillée de questions que pose la détermination des facteurs de rajustement des poids en vue de combiner deux panels de l'EDTR fondée sur la méthodologie présentée ici est décrite dans Latouche et coll. (2000). La question de la variance des estimations transversales n'est discutée dans le présent article que dans le contexte de l'estimation de la variance liée aux changements bimensuels dans l'échantillon au cours du temps, partiellement les mouvements d'une strate à l'autre, sont exposés dans Merkouris (1999). Notons, en guise de conclusion, que la qualité d'une méthode d'estimation transversale dépend de la possibilité de définir les divers

L'auteur remercie Milorad Kovacevic, Michel Latouche, Pierre Lavallée et Harold Maanel de leurs commentaires précieux. Les commentateurs et suggestions détaillées formulées par trois examinateurs au sujet d'une version antérieure de cet article ont permis d'en améliorer le contenu et la présentation.

BIBLIOGRAPHIE

BANKIER, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.

DEVILLE, J.C. (1998). Les enquêtes par panel : En quoi diffèrent-elles des autres enquêtes? Suivi de comment attraper une population en se servant d'une autre. *Actes des Journées de méthodologie statistiques*, numéro 84-85-86, 63-82.

KALTON, G., et ANDERSON, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65-82.

KALTON, G., et CITRO, C.F. (1993). Enquêtes par panel: Ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.

KALTON, G., et BRICK, J. M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-49.

LATOUCHE, M., DUFOUR, J. et MERKOURIS, T. (2000). Cross-sectional weighting for the SLID : Combining two or more panels. *Income Research Paper Series*, 75F0002MIE6, Statistique Canada.

LAVALLÉE, P. (1994). Ajout du second panel à l'EDTR : sélection et pondération. Document interne, Statistique Canada.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LAVIGNE, M., et MICHAUD, S. (1998). General aspects of the I'EDTR 98-05 E, Statistique Canada.

MERKOURIS, T. (1999). On the weight share method for panel household surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 255-260.

SINGH, A.C., et WU, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 69-77.

SKINNER, C.J., et RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

que $E(w_i) = 0$ si $M_i' \neq 0$, puisque $w_i \neq 0$ uniquement si $M_i' > 0$. Pour la caractéristique étudiée y , nous pouvons exprimer le total pour la population de personnes au temps t sous la forme $Y = \sum_{i=1}^H y_{ik}$, où y_{ik} représente la valeur de y pour la personne k dans le ménage \mathcal{H}_i . Alors, un estimateur de Y est donné par

$$\hat{Y} = \sum_{i=1}^H w_i \sum_{k=1}^{N_i} y_{ik} \quad (11)$$

$$= \sum_{i=1}^H w_i \left[\sum_{k=1}^{M_{B_i}} y_{ik} + \sum_{k=1}^{M_{A_i}} y_{ik} + \sum_{k=1}^{N_i - M_i'} y_{ik} \right]$$

$$= Y^B + Y^A + Y^C,$$

avec w_i défini comme dans (10), A^C représentant l'ensemble de personnes ne faisant pas partie de la base de sondage A et la notation évidente pour le deuxième membre de (11). Dans (11), l'estimateur \hat{Y} est donné par la somme de trois estimateurs, Y^B , Y^A et Y^C , pour les totaux ayant trait aux domaines de population B , a et A^C , respectivement. Les estimateurs Y^B et Y^A sont non biaisés, même s'ils sont fondés sur un ensemble d'unités qui ne sont pas nécessairement identiques aux échantillons originaux $s_{ab} \cup s_a$ et s_a , respectivement. Par exemple, l'estimateur Y^B se fonde sur un ensemble d'unités comprenant les unités restantes d'un échantillon combiné original s_{ab} provenant de la base de sondage B et, éventuellement, les cohabitants présents au départ dans B . L'estimateur Y^A n'est pas dépourvu de biais pour Y^A , car les personnes comprises dans A^C qui vivent dans les ménages ne contenant aucun membre de la population originale ne sont pas représentées dans l'enquête par panel. Néanmoins, l'estimateur Y^A est non biaisé pour le total correspondant au reste de A^C qui est représenté dans les panels combinés par les cohabitants absents au départ. Dans le cas spécial où le temps t coïncide avec le début du deuxième panel (ou avec le moment de la sélection d'un échantillon supplémentaire), $A^C = \emptyset$, $N_i' = M_i'$ et l'estimateur $Y = Y^B + Y^A$ est non biaisé pour Y . Notons ici que, si les poids appliqués aux personnes qui répondent au temps t sont corrigés pour la non-réponse, la relation $E(w_i) = 1$ pourrait n'être vérifiée qu'approximativement et, en ce sens, les estimateurs résultants pourraient n'être qu'approximativement non biaisés.

Il est important de souligner que, dans (11), l'estimateur \hat{Y} peut être exprimé sous la forme

$$\hat{Y} = \sum_{i=1}^I w_i' Y_i',$$

où $Y_i' = \sum_{k=1}^{N_i'} y_{ik}$ représente le total pour le ménage \mathcal{H}_i' . Donc, Y est également un estimateur du total au niveau du ménage au temps t .

Comme le rajustement des poids effectué lors de la combinaison des panels, le rajustement par la méthode du partage des poids peut être fait au niveau d'une superstrate,

disons la province, pour l'échantillon combiné de chaque province. Dans ces conditions, les personnes qui, au temps t , résident dans une autre province que celle dans laquelle elles résidaient au moment de la sélection de l'un des panels sont traitées comme absentes au départ, puisqu'elles n'étaient pas membres de la population originale de leur nouvelle province. En particulier, les personnes (sélectionnées ou non sélectionnées dans leur province originale) qui déménagent d'une province à une autre et qui font partie d'un ménage longitudinal dans leur nouvelle province au temps t sont traitées comme des cohabitants absents au départ. Si l'on utilise un échantillon de remise à niveau au temps t , les personnes qui déménagent d'une province à l'autre sont traitées comme des cohabitants présents au départ. L'application de la méthode du partage des poids séparément à chaque superstrate offre certains avantages opérationnels et statistiques par rapport à la méthode type de partage des poids. Merkouris (1999) a comparé les mérites respectifs des deux méthodes.

5. INTÉGRATION DES DIVERS RAJUSTEMENTS DES POIDS

En plus de ceux décrits jusqu'à présent, d'autres rajustements des poids appliqués aux données d'une enquête-ménage par panel peuvent être nécessaires. Voici une brève description de la série intégrée des divers rajustements de poids.

Le premier, appliqué au niveau des unités d'échantillon originales, tient compte de la non-réponse à un cycle, qui a lieu lorsqu'une unité échantillonnée participe à certains cycles auxquels elle est admissible, mais pas à tous. Pour une discussion du rajustement des poids pour la non-réponse à un cycle, consulter Kalton et Brick (1995). Le rajustement est fait séparément pour les divers panels de chaque cycle.

Le deuxième rajustement est celui qui permet de combiner les échantillons des divers panels en un échantillon unique pour produire des estimations transversales. Il s'applique aux poids des unités échantillonnées des panels, corrigés pour la non-réponse à un cycle, et se fait selon la méthode décrite à la section 3.

Le troisième rajustement comprend l'application de la méthode du partage des poids à l'échantillon combiné de panels lors de n'importe quel cycle réalisé après la création du panel le plus récent, selon la méthode décrite à la section 4.

Enfin, lors du rajustement des poids par calage, les poids appliqués aux unités du panel combiné sont rajustés de façon à ce que les totaux estimatifs calculés pour certaines caractéristiques auxiliaires soient égaux aux totaux connus de population pour ces caractéristiques au moment du cycle courant, ce qui, dans le cas simple de la figure 1, correspond aux totaux pour la base de sondage complète A . Dans des situations plus générales, après la sélection du panel le plus récent, les totaux de calage incluront les

est nécessaire à cause des changements qui surviennent dans la composition des ménages après la sélection des

panels.

La transversale des poids est une méthode de pondération transversale qui consiste à attribuer un poids de base à chaque membre d'un ménage sélectionné dans le panel lors de tout cycle de l'enquête réalisé après le premier. Plus précisément, la méthode du partage des poids, telle qu'appliquée à un panel unique, revient à attribuer un poids positif à des personnes non sélectionnées dans le panel qui se sont jointes aux ménages contenant au moins une personne sélectionnée dans l'échantillon original. À l'instar de Lavallée (1995), nous qualifions ici ce genre de ménages de ménages longitudinaux et nous qualifions de cohabitants les personnes non sélectionnées dans le panel, mais qui vivent dans un ménage longitudinal. Nous faisons en outre la distinction entre les cohabitants présents au départ, c'est-à-dire ceux qui faisaient partie de la population (échantillonnée) originale et les cohabitants absents au départ, c'est-à-dire les nouveaux arrivants dans la population. La méthode du partage des poids permet aussi de traiter des situations problématiques telles que les ménages non sélectionnés dans le panel qui ont été formés après le premier cycle par des membres de ménages différents sélectionnés au départ, ainsi que les personnes sélectionnées au départ qui ont subéquemment déménagé dans d'autres ménages longitudinaux. Pour une discussion détaillée de la méthode du partage des poids pour un panel unique, consulter Kalton et Brick (1995) et Lavallée (1995). Pour appliquer la méthode du partage des poids à une enquête à panels multiples, nous devons tenir compte des points qui suivent. Dans le cas de ces enquêtes, pour les panels combinés, la population originale correspond à l'union des populations couvertes par les divers panels au moment de leur sélection. Par conséquent, l'échantillon original comprend toutes les unités sélectionnées dans l'échantillon combiné de panels; donc, un cohabitant présent au départ est une personne qui répondait aux critères de sélection dans n'importe quel panel. Ainsi, selon cette approche, lors de tout cycle réalisé après la sélection du panel le plus récent, un cohabitant peut être classé dans la catégorie des cohabitants présents ou absents au départ en ce qui concerne l'échantillon combiné original, mais non en ce qui concerne chaque panel original. Notons que, lors du premier cycle d'un nouveau panel, ou lorsqu'on utilise un échantillon de remise à niveau, tous les cohabitants sont présents au départ. En revanche, l'application de la méthode du partage des poids séparément à chaque panel (avant leur combinaison) nécessiterait des renseignements plus précis sur l'admissibilité des cohabitants à la sélection dans chacun des panels, afin de pouvoir faire la distinction entre les cohabitants présents au départ et ceux absents au départ, et de définir le domaine temporel qui inclut chacun des cohabitants. Or, il est plus que probable que l'on ne dispose pas de ces renseignements. En outre, combiner les panels après avoir appliqué la méthode du partage des poids

Pretons comme point de départ les conditions d'enquête représentées à la figure 1, où deux panels se chevauchent au point dans le temps où débute l'utilisation du deuxième panel et supposons que, lors d'un cycle ultérieur (temps t), la population comptera N_t personnes, avec N_t personnes dans le ménage \mathcal{H}_t , $t = 1, \dots, H$ et $\sum N_t = N$. Représentons par M_t le nombre de personnes dans le ménage \mathcal{H}_t , avec temps t qui appartiennent à la population originale, avec M_t^{Bi} et M_t^{ai} personnes provenant du domaine original de la base de sondage B et du domaine non chevauchant de la base de sondage a , respectivement, de sorte que $M_t = M_t^{Bi} + M_t^{ai}$. Certains des nombres M_t^{Bi} , M_t^{ai} et $N_t - M_t^{Bi} - M_t^{ai}$, si les poids aléatoires appliqués aux personnes dans B et a sont ceux définis à la section 3.1 et les poids appliqués aux $N_t - M_t^{Bi} - M_t^{ai}$ cohabitants absents au départ dans \mathcal{H}_t sont identiquement égaux à zéro, la méthode du partage des poids permet de définir un poids commun pour toutes les personnes comprises dans \mathcal{H}_t (y compris les nouveaux membres) ayant la forme

$$w_t^i = \frac{1}{\sum_{k=1}^M w_t^{ik}}, \quad (10)$$

où w_t^{ik} représente le poids du k^e membre du ménage qui appartient à la population originale. Il est alors évident que $E(w_t^i) = 1$ pour chaque ménage pour lequel $M_t^i \neq 0$, tandis

Il se pourrait fort bien que le rajustement par le facteur $1 - p$ exclue les poids qui leur sont appliqués, mais ceci n'aurait aucun effet sur l'estimation transversale, à moins que les nouveaux ne fassent partie de la population étudiée. En outre, le rajustement par le facteur $1 - p$ des poids attribués aux nouveaux dans s_a a l'effet favorable de produire un poids commun pour le ménage. Le calage des poids de l'échantillon combiné sur des totaux connus de la population de la base de sondage complète A amoindrirait la sous-représentation du reste du domaine a , qui est formé principalement d'immigrants, mais un certain biais pourrait persister si les caractéristiques d'échantillon des membres de cette partie de la population sont fort différentes de celles des membres du domaine de population B . À moins que l'intervalle de temps entre la sélection des deux panels soit important, la taille de cette partie de la population est très petite, comparativement à l'ensemble de la population, et l'effet de biais éventuel sur les estimations des totaux globaux devrait être négligeable.

Maintenant, la valeur optimale de p (c'est-à-dire celle qui minimise la variance) dans (7) est donnée par

$$(8) \quad p'' = \frac{\text{Var}(Y_s^A) + \text{Var}(Y_s^B)}{\text{Var}(Y_s^A) + \text{Var}(Y_s^B)}.$$

Si nous ne tenons pas compte des corrections pour une population finie, nous pouvons montrer que (8) peut s'exprimer sous la forme

$$(9) \quad p'' = \frac{n_B d^A N_2^A S_2^A + n_A d^B N_2^B S_2^B}{n_B d^A N_2^A S_2^A + n_A d^B N_2^B S_2^B + n_B d^A + n_A d^B}.$$

avec $c = (N_2^B S_2^B) / (N_2^A S_2^A)$, et où n_B , n_A sont les tailles des échantillons s_B et s_A , d^A , d^B sont les effets de plan associés à s_B et s_A et la caractéristique Y , N_2^A , N_2^B sont les tailles des bases de sondage A et B et S_2^A , S_2^B sont les variances de la caractéristique Y dans A et B . Si nous tenons compte du fait que N_B pourrait n'être que légèrement plus petite que N_A (selon l'intervalle entre la sélection de deux panels) et que nous supposons que les variances inconnues S_2^A et S_2^B sont presque égales, nous obtenons une bonne approximation pratique de la valeur optimale de p en fixant simplement la valeur de c égale à l'unité dans (9). L'hypothèse voulant que les variances S_2^A et S_2^B soient presque égales est raisonnable, compte tenu de la grandeur de N_B comparativement à N_A . Nous pourrions utiliser des valeurs approximatives de d^A et d^B tirées d'autres enquêtes dont les plans de sondage sont les mêmes que ceux utilisés pour sélectionner les deux panels, de préférence pour une caractéristique comme la taille d'un grand domaine de population. Maintenant, si Y_c et Y_1 représentent l'estimateur Y_p^A dans (7) lorsqu'on

$$\frac{\text{Var}(Y_c) - \text{Var}(Y_1)}{\text{Var}(Y_c)} = - \frac{c}{(c-1)^2} \frac{p_1''}{(1-p_1'')}.$$

Pour une valeur de c s'approchant vraisemblablement de 1,0, la perte d'efficacité sera négligeable.

Il est intéressant de comparer l'efficacité de l'estimateur donné par (7), lorsqu'on choisit p'' comme en (8), comparativement à l'estimateur optimal donné par l'équation (3), lorsqu'on choisit p comme dans l'équation (4), utilisé quand le domaine s_a est définissable. Représentons ces estimateurs par Y_1^A et Y_1^B , respectivement. Alors, en nous servant de l'inégalité $\text{Cov}^2(Y_s^A, Y_s^B) \leq \text{Var}(Y_s^A) \text{Var}(Y_s^B)$, nous pouvons montrer que $\text{Var}(Y_1^A) - \text{Var}(Y_1^B) \geq (p'' - p') \text{Var}(Y_s^A)$, où p' est choisi comme dans (5). Comme nous l'avons déjà mentionné, généralement, $\text{Cov}(Y_s^A, Y_s^B) > 0$, si bien que $p'' > p'$ et, donc, $\text{Var}(Y_1^A) \geq \text{Var}(Y_1^B)$. Par conséquent, malgré l'utilisation des valeurs exactes de p'' et p' dans la comparaison, l'approche adoptée dans cette sous-section aboutirait, dans la plupart des cas, à une réduction de la variance des estimateurs calculés. La borne inférieure du gain d'efficacité comparativement à Y_1^A serait alors donnée par

$$\frac{\text{Var}(Y_1^A) - \text{Var}(Y_1^B)}{\text{Var}(Y_1^A)} > \frac{1 - p'}{1 - p''}.$$

L'extension de la procédure de rajustement des poids décrite plus haut aux enquêtes, comptant plus de deux panels dont les domaines temporels d'échantillon sont indépendants est simple. Nous aurons dans ce cas autant de facteurs de rajustement des poids, dont la valeur totale sera égale à l'unité, qu'il y aura de panels. Cette méthode très pratique produira de bonnes estimations transversales dans le cas d'enquêtes à panels multiples pour lesquelles l'intervalle entre la sélection des divers panels n'est pas grand. Sinon, le biais éventuellement causé par l'erreur de définition du domaine pourrait être préoccupant, principalement pour les estimations se rapportant à des sous-populations composées en proportion importante de nouveaux arrivants.

4. MÉTHODE DU PARTAGE DES POIDS POUR LES PANELS COMBINÉS

Nous décrivons dans cette section l'application d'une méthode de rajustement des poids, connue sous le nom de méthode de partage des poids, à l'échantillon combiné des panels obtenu pour n'importe quel cycle réalisé après la création du panel le plus récent. Ce rajustement des poids

les tailles réalisées lors de tout cycle de l'enquête, corrigées pour tenir compte des effets de plan de sondage) peut être fort différentes si les taux d'érosion et les effets de plans de sondage diffèrent pour les deux panels. En outre, si les tailles des domaines d'échantillon s_{ab} et s_a sont à peu près proportionnelles aux tailles des domaines de population correspondants, $\text{Var}(Y_{s_a}^A)$ sera nettement, disons k fois, plus petit que $\text{Var}(Y_{s_{ab}}^A)$. Alors,

$$2 \text{Cov}(Y_{s_{ab}}^A, Y_{s_a}^A) \leq 2 \sqrt{\text{Var}(Y_{s_{ab}}^A) \text{Var}(Y_{s_a}^A)}$$

$$= 2 \frac{\sqrt{k}}{\text{Var}(Y_{s_{ab}}^A)},$$

si bien qu'une condition suffisante pour que l'estimateur Y_P^A soit plus efficace que l'estimateur « de triage » est

$$2 \frac{\sqrt{k}}{\text{Var}(Y_{s_{ab}}^A)} > \text{Var}(Y_{s_b}^A).$$

Cette condition signifie que, si $\text{Var}(Y_{s_b}^A)$ n'est pas très petit comparativement à $\text{Var}(Y_{s_{ab}}^A)$, on ne peut ignorer le domaine d'échantillon s_{ab} lorsqu'on estime Y_A . Elle est ordinairement satisfaite dans le cas des enquêtes-ménages par panel. Un autre argument en faveur de l'inclusion de s_{ab} dans l'estimation est que la qualité de ce domaine est susceptible d'être biaisé par l'érosion de l'échantillon.

Le facteur approximatif simple de rajustement des poids p' donné par l'expression (6) permet de réaliser une combinaison efficace d'échantillons de panel, tenant compte de la précision de $Y_{s_b}^A$ comparativement à $Y_{s_{ab}}^A$ grâce aux tailles effectives d'échantillon n_b/d_b et n_{ab}/d_{ab} . Ces tailles effectives d'échantillon varient en fonction du temps, mais leur ratio (et donc p') devrait être assez stable durant la période de chevauchement des panels. En ce qui concerne le calcul de la variance, puisque n_{ab} est ordinairement quasi non aléatoire, il est possible et commode de traiter le facteur de rajustement p' comme une constante dans toute méthode d'estimation de la variance.

Il importe de souligner ici que des gains supplémentaires d'efficacité pourront être réalisés par l'intégration de données auxiliaires dans les poids grâce à leur rajustement par calage sur des totaux de population connus. Enfin, mentionnons que, si le critère sur lequel se fonde le choix de la valeur de p est la minimisation de l'erreur quadratique moyenne de la composante à base commune $Y_P^B = (1-p)Y_{s_b}^B$ de l'estimateur Y_P^A , alors nous pouvons montrer facilement que, si $Y_{s_b}^B$ et $Y_{s_{ab}}^B$ sont entachés d'un même biais, la valeur optimale de p est la même que celle donnée par l'équation (5). Néanmoins, nous ne nous attendons pas à ce que les biais soient égaux; par exemple, l'écart entre les taux d'érosion de l'échantillon observés pour les deux panels pourrait produire des niveaux

Généralisation aux enquêtes à panels multiples et discussion d'autres méthodes

La méthode de pondération décrite plus haut s'applique à la situation simple d'une enquête à deux panels au moment de la création du deuxième panel. Lors de cycles ultérieurs de l'enquête, un domaine de base de sondage non chevauchant supplémentaire, que nous représentons par b , peut être formé par le retour dans la base de sondage B des personnes qui l'avaient quittée. Les unités provenant de b sélectionnées au départ dans le premier panel n'étaient pas présentes lors de la sélection du deuxième. Donc, maintenant, nous ne devons pas rajuster les poids dans le domaine d'échantillon non chevauchant s_b en vue de combiner les deux panels. Qui plus est, la valeur de p ne sera pas affectée, puisqu'elle est fondée uniquement sur le domaine chevauchant de l'échantillon combiné. Tout comme le fait de ne pas tenir compte du domaine d'échantillon s_a pour déterminer la valeur de p , ignorer le domaine d'échantillon s_b beaucoup plus petit et éventuellement vide aura un effet négligeable sur l'efficacité des estimateurs dérivés.

Étant donné sa simplicité, la méthode de pondération proposée pour combiner les deux panels peut être généralisée sans difficulté au cas des enquêtes comportant plus de deux panels chevauchants. En pratique, la généralisation la plus probable porterait sur trois panels. La construction d'un échantillon transversal combiné nécessiterait alors le rajustement des poids d'échantillonnage des unités provenant des domaines temporels des divers panels qui représentent un domaine temporel commun de la population. Pour chaque domaine temporel commun de population, les facteurs de rajustement des poids seront fondés sur les tailles d'échantillon effectives relatives des domaines de panel correspondants, par analogie avec l'expression (6) et leur somme sera égale à l'unité. Le nombre de domaines temporels de base de sondage et, donc, le nombre d'ensembles indépendants correspondants de facteurs de rajustement, sera assez petit, à cause du degré élevé d'emboîtement de la série de bases de sondage de panel. Par exemple, pour une enquête à trois panels, nous aurons un ensemble de trois facteurs de rajustement et un ensemble de deux.

Si nous faisons maintenant un retour en arrière, nous pouvons spécifier divers facteurs de rajustement des poids à un niveau inférieur de regroupement d'échantillons, comme un niveau de stratification particulier. Pour des

Notons que la condition $E(w_i^*) = 1$ est également nécessaire pour que la méthode du partage des poids soit valide lorsqu'on l'applique à l'échantillon combiné s pour tout cycle d'enquête réalisée après la sélection du deuxième panel. Un autre choix de la valeur de p se fonde sur la minimisation de la variance de la composante A à base de sondage commune $X_B^p = D^p + (1 - D^p)X_{s_{ab}}^p$ de l'estimateur X_A^p dans (3). Cette minimisation limitée, qui ne tient pas compte de l'estimateur par domaine X_{s_a} , ordinairement petit, donne la valeur

$$(5) \quad p' = \frac{\text{Var}(X_{s_{ab}}^p)}{\text{Var}(X_{s_a}^p) + \text{Var}(X_{s_{ab}}^p)},$$

qui est indépendante du terme de covariance et toujours comprise en zéro et un. Si nous minimisons la variance de X_B^p en imposant comme condition la valeur réalisée de la taille aléatoire n_{ab} du domaine d'échantillon s_{ab} , puis que nous utilisons la formule bien connue de la variance de l'estimateur d'un total dans des conditions d'échantillonnage aléatoire simple et que nous ignorons les corrections pour une population finie, nous pouvons montrer que (5) peut être calculé approximativement par

$$(6) \quad p' = \frac{n_B/p_B + n^{ab}/p^{ab}}{n_B/p_B},$$

où n_B est la taille de l'échantillon s_B , et d_B , d_{ab} sont les effets de plan de sondage associés à s_B et s_{ab} . Le calcul de la valeur de p' nécessite les estimations des deux effets de plan de sondage qui doivent se fonder sur s_B et s_{ab} . Des valeurs approximatives approchées de d_B et d_{ab} peuvent être tirées d'autres enquêtes ayant les mêmes plans d'échantillonnage que les deux panels. Cependant, comme p' dépend de la caractéristique y par la voie de d_B et d_{ab} , nous devons calculer un ensemble différent de poids pour chaque caractéristique étudiée. Outre le fait qu'ils rendent le processus d'estimation opérationnellement peu commode, les ensembles distincts de poids peuvent produire des estimations incohérentes. L'obtention de valeurs approchées de d_B et d_{ab} de préférence pour une variable de dénombrement associée à un grand domaine de population, représente un compromis. La méthode d'estimation proposée par Skinner et Rao (1996) dans le cas d'enquêtes à bases de sondage doubles comprend implicitement le même compromis. Notons que, puisque p' dépend de la caractéristique y uniquement par la voie du rapport d_B/d_{ab} , la perte d'efficacité des estimateurs des totaux ne devrait pas être importante pour d'autres caractéristiques. Soulignons aussi que, étant donné le décalage temporel entre la sélection des deux panels, les effets de plan de sondage seront différents et se manifesteront donc dans (6), même si les plans d'échantillonnage sont les mêmes pour les deux panels. Si nous utilisons des estimations des effets de plan de sondage fondés sur des données externes, le caractère aléatoire de p' tient uniquement à la taille

$$(5), \text{ respectivement. Alors, un calcul simple donne}$$

$$\frac{\text{Var}(X_{s_a}^p) - \text{Var}(X_{s_{ab}}^p)}{\text{Cov}(X_{s_{ab}}^p, X_{s_a}^p)} = \frac{\text{Var}(X_{s_a}^p) + \text{Var}(X_{s_{ab}}^p)}{\text{Var}(X_{s_{ab}}^p) \text{Var}(X_{s_a}^p)} \leq p' = \frac{\text{Var}(X_{s_a}^p)}{\text{Var}(X_{s_{ab}}^p)},$$

de sorte que la borne supérieure de la perte d'efficacité peut être exprimée comme suit

$$\frac{\text{Var}(X_{s_a}^p) - \text{Var}(X_{s_{ab}}^p)}{\text{Var}(X_{s_{ab}}^p)} \leq p' \frac{\text{Var}(X_{s_a}^p)}{\text{Var}(X_{s_{ab}}^p)}.$$

Étant donné la taille habituellement très petite de $X_{s_a}^p$ comparativement à $X_{s_{ab}}^p$ (la taille du domaine a est approximativement le quartième de la taille de la base de sondage complète A dans le cas de l'EDTR), il semble que la perte d'efficacité sera très faible pour la plupart des enquêtes-ménages par panel.

Il est intéressant de se demander si $X_{s_a}^p$ est ou non plus efficace que l'estimateur « de triage » simple $X_A^p = Y_{s_a} + Y_{s_{ab}}$, dont la variance est $\text{Var}(X_A^p) = \text{Var}(Y_{s_a}) + \text{Var}(Y_{s_{ab}})$ montrant aisément que $\text{Var}(X_{s_a}^p) > \text{Var}(X_A^p)$. Cette condition est certainement vérifiée si la covariance de $X_{s_a}^p$ et $X_{s_{ab}}^p$ est négative, ce qui pourrait être le cas si la valeur de la caractéristique estimée n'est pas la même pour les immigrants que pour les non-immigrants. En général, cette covariance peut, effectivement, être positive, car $X_{s_a}^p$ et $X_{s_{ab}}^p$ sont fondés sur les mêmes grappes aréolaires sélectionnées. Toutefois, dans ce cas aussi, la condition sera fort probablement vérifiée, étant donné la grandeur de $\text{Var}(X_{s_a}^p)$ comparativement à $\text{Var}(X_{s_{ab}}^p)$, et celle de $\text{Var}(X_{s_a}^p)$ comparativement à $\text{Var}(X_{s_{ab}}^p)$. En effet, typiquement, en vertu du plan de sondage, les échantillons de panel s_B et s_{ab} sont de même taille, alors que les tailles effectives des panels (c'est-à-dire

$X_B = \sum_B w_i^* y_i$, du total $X_B = \sum_B y_i$, quelles que soient les valeurs choisies des constantes p_i , qui satisfont $0 \leq p_i \leq 1$, et pour tous plans d'échantillonnage $p_A(s_A)$ et $p_B(s_B)$. L'équation (2) peut aussi être écrite sous la forme $w_i^* = p_i w_{Bi} + (1 - p_i) w_{Ai}$, où la définition des poids w_{Bi} et w_{Ai} est associée de façon évidente aux échantillons s_B et s_A . Donc, la catégorie de scénarios de pondération définie par l'équation (2) représente essentiellement des combinaisons pondérées différentes des poids dans les échantillons originiaux s_B et s_A . Les limites fixées pour les valeurs de p_i assurent que le poids w_i^* ne soit pas négatif. Notons que le poids dont le calcul est impossible $w_i^* = (\pi_{Ai}^* + \pi_{Bi}^*)^{-1}$ pour $i \in s \cap B$, utilisé dans (1) est un cas spécial de w_i^* avec $p_i = \pi_{Bi}^* (\pi_{Ai}^* + \pi_{Bi}^*)^{-1}$. Manifestement, le scénario de pondération défini par (2) ne permet pas d'éliminer les unités d'échantillonnage en double qui se retrouvent dans les deux échantillons. Si l'on impose la contrainte opérationnelle consistant à exclure de s_A les personnes déjà sélectionnées dans s_B , le deuxième terme du deuxième membre de l'équation (2) doit être modifié pour devenir $(1 - p_i) [\pi_{Ai}^* / (1 - \pi_{Bi}^*)]^{-1}$, afin d'assurer que $E(w_i^*) = 1$. Toutefois, $I\{i \in s_{ab}, i \in s_B\}$, afin d'assurer que $E(w_i^*) = 1$. Toutefois, cette modification pourrait être impossible, parce que, pour la réaliser, il faut connaître les probabilités d'inclusion des unités d'échantillonnage dans chacune des deux bases de sondage. Notons aussi que, si l'on impose la condition que les unités en double soient exclues, les deux échantillons ne seront plus indépendants. Néanmoins, puisque nous supposons que les deux probabilités π_{Ai} et π_{Bi} sont faibles, la probabilité qu'il existe des unités en double sera négligeable et, donc, tout biais dû à l'utilisation des scénarios de pondération résolubles définis par (2) seraient également négligeables. Si nous émettons cette hypothèse, nous devrions interpréter les deux variables indicatrices qui figurent dans (2) comme satisfaisant $I\{i \in s_B\} I\{i \in s_{ab}\} = 0$.

La question qui se pose maintenant est celle de choisir la valeur optimale de p_i , pour toute unité $i \in s \cap B$, conformément à certains critères de pondération optimale pour l'échantillon combiné. Une méthode consiste à choisir la valeur de p_i de façon à minimiser la variance du total $X^A = \sum_B w_i^* y_i + \sum_A y_i = (\pi_{Ai}^*)^{-1} I\{i \in s\}$ pour $i \in A$. Cependant, la minimisation de la variance de X^A par rapport à p_i pour toutes les unités $i \in s \cap B$ est insoluble. Une option plus simple consiste à limiter la catégorie de scénarios de pondération définis par l'équation (2) à un seul scénario dans lequel les facteurs d'ajustement des poids sont précisés non pas au niveau de l'unité, mais à un niveau d'agrégation plus élevé, qui pourrait correspondre à une strate ou à la base de sondage chevauchante complète B . Nous discutons de façon plus approfondie du niveau d'ajustement à la fin de cette sous-section. Il suffit, pour le développement de la méthode de pondération, de considérer le cas suivant, qui correspond à l'application d'un facteur unitaire d'ajustement des poids p pour la base de sondage B toute entière.

Détermination de la valeur de p . Problèmes pratiques et questions d'efficacité.

Si l'on applique un facteur uniforme de rajustement des poids p , la catégorie de scénarios de pondération définis par l'équation (2) pour la base de sondage B génère, pour le total global X_A , un groupe d'estimateurs non biaisés de la

$$(3) \quad Y_p^A = p Y_{s_B}^A + (1 - p) Y_{s_{ab}}^A + Y_{s_A}^A$$

où $Y_{s_A}^A$ et $Y_{s_{ab}}^A$ sont des estimateurs d'Horvitz-Thompson indépendants de $Y_{s_B}^A$ fondés sur s_B et s_{ab} , respectivement et $Y_{s_{ab}}^A$ est l'estimateur d'Horvitz-Thompson de Y_a fondé sur s_{ab} . Les valeurs limites de p donnent lieu à deux cas spéciaux de l'estimateur Y_p^A , dans chacun desquels le total du domaine chevauchant Y_B est estimé à partir d'un seul panel. Lorsqu'on donne une valeur nulle à p dans l'équation (3), l'estimateur trivial résultant Y_p^A pour l'ensemble de la population se fonde uniquement sur s_A . Le cas où l'on donne à p la valeur de un dans l'équation (3) est plus intéressant. Le simple estimateur non biaisé $Y^A = Y_{s_A}^A + Y_{s_{ab}}^A$ qu'il sous-entend serait l'estimateur naturel dans le cas d'une enquête par panel comportant un panel et un échantillon transversal supplémentaire dont les unités ont été « triées » de façon à ne dénombrer que celles faisant partie du domaine des nouveaux arrivants. Dans de telles conditions, cet estimateur simple représenterait un cas spécial d'un estimateur « de triage » à bases de sondage multiples dont la caractéristique particulière est la nature temporelle du domaine de la base non chevauchante a . Dans le présent contexte, l'estimateur de triage semble inefficace, car l'information qui figure dans le domaine d'échantillon s_{ab} n'est pas utilisée. Nous pourrions mieux exploiter les données provenant des deux panels en combinant s_B et s_{ab} , au moyen d'un facteur p optimal fondé sur la minimisation de la variance de Y_p^A . La valeur optimale de p est donnée par

$$(4) \quad p = \frac{\text{Var}(Y_{s_{ab}}^A) + \text{Cov}(Y_{s_{ab}}^A, Y_{s_A}^A)}{\text{Var}(Y_{s_{ab}}^A) + \text{Var}(Y_{s_A}^A)}$$

Nous ne connaissons pas les termes de variance et de covariance qui figurent dans (4), mais nous pourrions les estimer d'après les données d'échantillon, auquel cas la valeur choisie de p réduirait effectivement au minimum la variance estimée de Y_p^A . Le choix de cette valeur de p présente de nombreux inconvénients. En général, l'estimation de la valeur optimale de p est difficile; en cas d'enquête comportant plus de deux panels, il serait plus commode d'estimer l'ensemble requis de ces rajustements de poids. En outre, une estimation d'échantillon de la valeur optimale de p dans (4) augmente la variabilité de l'estimateur Y_p^A et complique l'estimation de sa variance. De surcroît, le fait que la valeur optimale estimée de p dépende de données d'échantillon sous-entend que $E(w_i^*) \neq 1$ pour $i \in B$, ce qui affaiblit le caractère non biaisé de l'estimateur (3).

2.2 Analogie avec une enquête à bases de sondage multiples

Compte tenu de ce qui précède, nous pouvons considérer une enquête à panels multiples chevauchants comme un cas spécial d'enquête à bases de sondage multiples où la base de sondage pour la population transversale correspond à l'union de domaines temporels s'excluant mutuellement qui sont définis par les bases de sondage des panels et de leurs intersections. La taille des bases de sondage des panels interseptions. La taille des bases de sondage des panels individuels et les caractéristiques des membres de la population de la base de sondage de chaque panel varient au fil du temps. Cette propriété contraste avec le système qui caractérise les enquêtes à bases de sondage multiples habituelles. De surcroît, le degré d'embordement de la série de bases de sondage des panels est élevé, si bien que le nombre total de domaines temporels qui s'excluent mutuellement est faible. Parmi les divers domaines de base de sondage, celui qui est commun à tous les panels est de loin le plus grand. Comme nous l'expliquons à la section suivante, ces caractéristiques particulières des bases de sondage multiples influent sur l'estimation transversale.

Les domaines temporels d'échantillon pourraient être encore moins statiques à cause de l'érosion, des mouvements de certaines personnes dans les panels et entre ceux-ci, et des mouvements des membres non sélectionnés des ménages dans lesquels vivent les membres sélectionnés du panel. Par exemple, si de nouvelles personnes (comme des immigrants) arrivent dans les ménages où vivent des membres d'un panel particulier, ce dernier chevauchera la limite entre sa propre base de sondage et celle du panel suivant.

L'analogie avec les échantillons d'enquête à bases de sondage multiples nous permet d'étudier dans un cadre familier le problème de l'estimation transversale dans le cas d'enquêtes répétées à panels chevauchants. Cependant, nous devons tenir compte des caractéristiques dynamiques distinctives des enquêtes à panels multiples si nous voulons nous inspirer des méthodes classiques appliquées aux enquêtes à bases de sondage multiples pour élaborer une méthode d'estimation transversale.

Pour introduire une méthode d'estimation transversale qui combine les données recueillies auprès des divers panels d'une enquête-ménage par panel répétée, il suffit de considérer le cas simple de deux panels chevauchants au moment de la création du deuxième panel. Notons que cette situation serait celle observée systématiquement dans le cas d'une enquête comportant un panel et un échantillon de remise à niveau. Par conséquent, si nous adoptons la notation classique appliquée aux bases de sondage multiples, et que nous représentons par B et A les bases de sondage des premier et deuxième panels ($B \subset A$) au moment de la création du deuxième, et par s_B, s_A les échantillons respectifs, nous pouvons représenter schématiquement la situation par le diagramme de la figure 1.

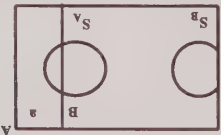


Figure 1. Deux panels chevauchants au moment de la création du deuxième.

Dans la figure 1, A représente la base de sondage complète, si bien qu'au moment de sa création, le deuxième panel représente la population transversale à ce point précis dans le temps. Le domaine chevauchant B correspond au reste de la base de sondage originale du premier panel. Le domaine $a = B \cap A$ comprend toutes les nouvelles personnes qui sont entrées dans la population depuis la création du premier panel. Les échantillons s_B et s_A sont les échantillons sélectionnés au départ, où la taille de s_B a été réduite pour tenir compte des personnes qui ont quitté la population et des non-répondants. Nous supposons que les échantillons s_A et s_B sont sélectionnés indépendamment à partir de A et B conformément aux plans de sondage probabilistes précisés $p_A(s_A)$ et $p_B(s_B)$, qui déterminent les probabilités d'inclusion π_A^i et π_B^i , et de la i^{e} unité (ménage ou personne dans ce ménage) pour les échantillons originaux s_A et s_B , respectivement. Les échantillons s_A et s_B peuvent se recouper, puisque les personnes comprises dans la base de sondage chevauchante B peuvent être sélectionnées dans les deux panels. La question du chevauchement des panels (échantillons) est semblable à celle des unités d'échantillon-nage en double que posent les enquêtes à bases de sondage multiples. Dans le cas d'enquêtes-ménages par panel répétées, pour alléger le fardeau de réponse, nous pourrions adopter une contrainte opérationnelle consistant à exclure de s_A les personnes déjà sélectionnées dans s_B , donc d'induire $s_A \cap s_B = \emptyset$ pour une discussion à ce sujet, consulter Lavalée (1994). Ici, comme dans tout cas d'enquête à bases de sondage multiples, nous constatons que les valeurs de π_A^i et π_B^i sont faibles et que la probabilité de sélectionner des unités en double est négligeable. Nous poserons dans la suite que les probabilités π_A^i et π_B^i sont faibles et, en fait, que $s_A \cap s_B = \emptyset$.

3. PONDERATION ET ESTIMATION TRANSVERSALES

Dans cette section, nous décrivons les méthodes qui permettent de combiner les données provenant de plusieurs panels d'une enquête-ménage par panel répétée en vue de produire des estimations transversales des paramètres de population. Nous limitons la discussion à l'estimation des totaux. Nous présentons une méthode uniformisée de calcul des estimations transversales, applicable aux ménages et aux personnes, qui se fonde sur la production d'un ensemble de poids pour l'échantillon résultant de la

(que nous appellerons par conséquent échantillon de remise à niveau) ne doit être utilisé qu'une seule fois, pour produire des estimations transversales, et, en principe, son effectif devrait être plus faible que celui du panel. Dans le contexte de la création d'un échantillon transversal, nous considérons l'échantillon de remise à niveau comme un cas non trivial d'échantillon supplémentaire, que nous traitons essentiellement comme un petit panel chevauchant additionnel.

Pour ce qui est des personnes qui quittent la population, la situation est la suivante. Pour tout panel, la base de sondage de la population observée au temps t est essentiellement celle de la population telle qu'elle était au moment de la création du panel, dans laquelle les personnes qui quittent la population dans l'intervalle sont représentées par des blancs. Les membres du panel qui quittent la population avant le temps t correspondent à des blancs dans la base de sondage et, donc, leur départ rend les estimations transversales au temps t moins efficaces, mais ne les biaisent pas; pour une discussion pertinente, consulter Kalton et Brick (1995).

Les observations qui précèdent nous mènent à voir comme suit la couverture de la population par les divers panels lors d'un cycle donné de l'enquête. Du point de vue transversal, au moment de sa sélection, chaque panel couvre entièrement la population étudiée représentée par les panels précédents. Par conséquent, les bases de sondage des panels forment une série chronologique, dans laquelle la base de sondage de chaque panel contient, au moment de la création de celui-ci, les bases de sondage des panels précédents. Dans le cas d'une série de bases de sondage de cette sorte, il y a formation séquentielle d'une base de sondage commune qui correspond à l'intersection de la base de sondage du nouveau panel avec le reste de la base de sondage commune originale des panels actifs précédents. Lors de tout cycle de l'enquête, la base de sondage commune correspond à celle qui existe au moment de la création du plus récent de ces panels, à l'exclusion de la population qui quitte la population. Au moment de la création d'un nouveau panel, le domaine de la base de sondage non chevauchante comprend les personnes qui se sont jointes à la population après la création du panel précédent. On peut aussi former d'autres domaines de base de sondage (de taille comparative très petite) en réintégrant des unités provenant d'anciennes bases de sondage, mais alors, la série chronologique des bases de sondage n'est pas complètement emboîtée. Étant donné cette dernière catégorie de domaines de base de sondage, lors de tout cycle de l'enquête, après la sélection du panel le plus récent, la base de sondage complète correspond à l'union des bases de sondage de tous les panels à ce point précis dans le temps, plutôt que simplement au reste de la base de sondage du panel le plus récent. Dans le cas des enquêtes par panel où l'on emploie un échantillon de remise à niveau lors de chaque cycle, la base de sondage complète est celle de l'échantillon de remise à niveau.

Le présent article décrit des méthodes d'estimations transversales visant à combiner les données provenant des panels chevauchants d'une enquête-ménage par panel répétée. À la section 2, nous discutons de la couverture de la population par les panels individuels lors d'un cycle donné et de l'utilisation des panels combinés, complétés par un échantillon de « remise à niveau », en vue de produire un échantillon transversal représentatif. Nous discutons aussi dans cette section des analogies avec une enquête à bases de sondage multiples et de la dynamique des problèmes de l'échantillon. À la section 3, nous décrivons les problèmes que posent la pondération et l'estimation dans le cas d'une enquête-ménage par panel répétée. Puis nous proposons des stratégies de pondération appropriées pour diverses formes d'enquête par panel. Nous discutons des problèmes de biais et d'efficacité que pose la combinaison des panels. À la section 4, nous décrivons une méthode de rajustement des poids qui tient bien compte des changements qui surviennent au cours du temps dans les panels combinés. À la section 5, nous parlons de l'intégration des divers rajustements des poids nécessaires pour produire des estimations transversales à partir des données d'une enquête-ménage par panel répétée. Enfin, à la section 6, nous résumons nos observations et présentons nos conclusions.

2. CONSIDÉRATIONS GÉNÉRALES

2.1 Couverture de la population transversale

Les changements de composition de la population qui surviennent au fil du temps, lorsque des personnes quittent la population ou s'y joignent, sont un élément important dont il faut tenir compte lors de l'estimation transversale. Dans le cas d'une enquête-ménage à panel unique, les nouveaux arrivants qui se sont joints à la population d'enquête depuis la création du panel ne sont pas représentés dans l'échantillon lors des cycles ultérieurs s'ils vivent dans des ménages qui ne comptent aucun membre de la population originale. Une enquête-ménage à panels multiples chevauchants offre une meilleure couverture de la population observée qu'une enquête à panel unique, car elle permet de réduire la durée de la période qui n'est couverte par aucun panel. Dans le cas de l'EDTR, cette période est réduite d'un maximum de six ans à un maximum de trois ans. Néanmoins, le problème de la couverture complète de la population persiste à moins que l'on ne sélectionne un échantillon supplémentaire spécial de la population non couverte lors de chaque cycle de l'enquête. Un scénario d'enquête comportant un panel et un échantillon supplémentaire sélectionné lors de chaque cycle de l'enquête en vue de produire des estimations transversales est décrit par Lavallée (1995). Une autre stratégie consiste à sélectionner, lors de chaque cycle, un nouvel échantillon qui couvre la population observée complète, mais ne constitue pas un nouveau panel. Cet échantillon

Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples

TAKIS MERKOURIS¹

RÉSUMÉ

Le présent article décrit les méthodes de pondération qui permettent de combiner les renseignements provenant de plusieurs panels d'une enquête-ménage par panel répétée en vue de produire des estimations transversales. Nous examinons la nature non statique d'une enquête par panel répétée dans le contexte de l'estimation des paramètres de population lors de tout cycle de l'enquête. Nous décrivons une enquête par panel répétée avec panels chevauchants comme étant un cas particulier d'enquête à bases de sondage multiples pour laquelle les bases de sondage des panels forment une série chronologique. Nous proposons des stratégies de pondération appropriées pour diverses enquêtes à panels multiples. Les scénarios de pondération proposés comprennent un rajustement des poids pour les domaines de l'échantillon combiné de panels qui correspondent à des périodes identiques couvertes par les panels individuels. Nous discutons d'une méthode de rajustement des poids qui tient compte des changements survenus dans les panels. Enfin, nous parlons de l'intégration des divers rajustements des poids nécessaires pour produire des estimations transversales dans le cas d'une enquête-ménage par panel répétée.

MOTS CLÉS : Enquêtes par panel répétées; bases de sondage multiples; domaines temporels; panels combinés; pondération transversale; méthode du partage des poids.

1. INTRODUCTION

Une enquête par panel permet de recueillir des données d'enquête sur les mêmes éléments d'un échantillon à diverses périodes (les cycles de l'enquête). Une enquête par panel répétée comprend une série d'enquêtes par panel, chacune de durée fixe, dont les panels respectifs sont sélectionnés à divers points dans le temps. Dans le cas d'une enquête-ménage par panel répétée, on sélectionne, pour former chaque panel, un échantillon de ménages représentatif de la population de ménages existants au moment de la création du panel. Selon les objectifs de l'enquête, un membre seulement de chaque ménage échantillonné ou tous font partie du panel et sont suivis pendant la durée du panel ou jusqu'à ce qu'ils ne fassent plus partie de la population observée. Lors d'un cycle subséquent de l'enquête, l'échantillon de ménages comprend tous les ménages dans lesquels vivent des membres du panel. Kalton et Citro (1993) ont passé en revue les diverses catégories d'enquêtes par panel. Deville (1998), quant à lui, offre une représentation formelle des concepts pertinents.

Les enquêtes-ménages par panel répétées que nous examinons ici comprennent au moins deux panels couvrant des périodes chevauchantes. L'Enquête sur la dynamique du travail et du revenu (EDTR) du Canada, réalisée auprès de deux panels chevauchants ayant chacun une durée de vie de six ans, en est un exemple type. Pour une description de l'EDTR, consulter Lavigne et Michaud (1998). Dans l'EDTR, chaque nouveau panel est créé trois ans après l'introduction du précédent. Dans le cas de chaque panel, l'échantillon est formé de deux groupes de renouvellement

provenant de l'Enquête sur la population active du Canada, réalisée auprès d'un échantillon stratifié à plusieurs degrés sélectionné à partir d'une base aréolaire dans laquelle les logements contenant les ménages sont les unités finales d'échantillonnage.

Quoique réalisée principalement en vue de recueillir des données longitudinales, une enquête par panel peut aussi servir à la production d'estimations transversales des paramètres de population lors d'un cycle donné. Pour produire les estimations transversales, des données sont habituellement recueillies lors de chaque cycle de l'enquête sur toutes les personnes vivant dans les ménages qui contiennent au moins un membre du panel longitudinal. La production d'estimations transversales par panel après le premier cycle pose des difficultés inhérentes à la dynamique de la population et du panel. Divers auteurs ont discuté des scénarios de pondération qui permettent de tenir compte des caractéristiques dynamiques d'un panel particulier, comme les personnes qui déménagent et les « cohabitants »; pour plus de précisions, consulter Kalton et Brick (1995), ainsi que Lavallée (1995). Pourtant, rares sont les travaux qui ont été publiés sur l'estimation transversale dans le cas d'une enquête-ménage par panel répétée avec panels chevauchants; certains travaux préliminaires réalisés dans le contexte de l'EDTR sont décrits par Lavallée (1994). Le problème qu'il faut résoudre pour produire des estimations transversales d'après des données d'enquête à panels multiples est celui de savoir quelle combinaison des panels rendrait compte comme il convient des changements qui surviennent dans la population et dans les panels au fil du temps.

¹ Takis Merkouris, Statistique Canada, Division des méthodes des enquêtes auprès des ménages, Parc Tunney, Ottawa (Ontario), K1A 0T6.

Comparativement aux méthodes 2 et 4, la méthode 3 est préférable dans certains cas. Comme il serait difficile, en pratique, de déterminer d'avance si c'est la méthode 3 ou la méthode 4 qui produira les coefficients de variation les plus faibles et que la méthode 3 peut produire d'importantes erreurs de couplage, la méthode 4 est celle qui devrait avoir la préférence. Par conséquent, la méthode classique d'utilisation de la MGPF avec la variable indicatrice I et la détermination des appariements par application d'une règle de décision telle que (2.3) semble la méthode la plus appropriée d'estimation du total Y^B d'après un échantillon sélectionné à partir de U^A .

REMERCIEMENTS

Les auteurs remercient le rédacteur adjoint et les deux examinateurs de leurs suggestions et commentaires utiles qui ont permis d'améliorer considérablement la qualité de l'article.

BIBLIOGRAPHIE

BARTLETT, S., KREWSKI, D., WANG, Y. et ZIELINSKI, J.M. (1993). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisés. *Techniques d'enquête*, 19, 3-13.

BELIN, T.R. (1993). Évaluation des sources de variation dans le couplage d'enregistrements au moyen d'une expérience factorielle. *Techniques d'enquête*, 19, 15-33.

BUPD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.

BUPD, E.C., et RADNER, D.B. (1969). The OBE size distributions series: Methods and tentative results for 1964. *American Economic Review*, Papers and Proceedings, LIX, 435-449.

ERNST, L. (1989). Weighing issues for longitudinal household and family estimates. Dans *Panel Surveys*, (Éd. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley and Sons, 135-159.

FELLEGI, I.P., et SUNTER, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

GAILLY, B., et LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg.

KALTON, G., et BRICK, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-49.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LIM, A. (2000). Results of the Linkage between the 1998 Taxation Data and the 1998 Farm Register. Document interne de la DMBE, Statistique Canada.

LYNCH, B.T., et ARENDS, W.L. (1977). Selection of a Summary Coding Procedure for the SRS Record Linkage System. Document of the Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

NEWCOME, H.B., KENNEDY, J.M., AXFORD, S.J. et JAMES, A.P. (1959). Automatic linkage of vital records. *Science*, 130, 954-959.

OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.

SÄRNDAHL, C.-E., SWENSSON, B. et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, A.C., MANTTEL, A.J., KINACK, M.D. et ROWE, G. (1993). Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle. *Techniques d'enquête*, 19, 67-89.

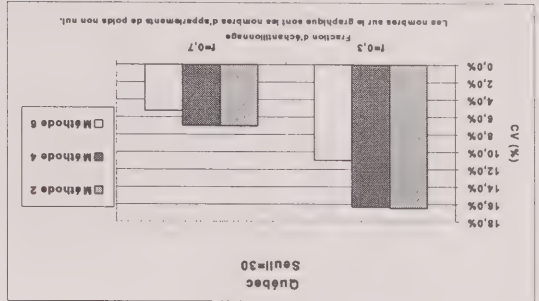
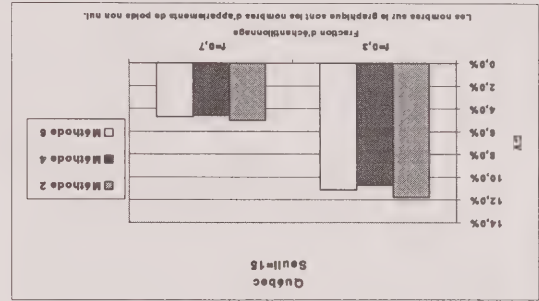
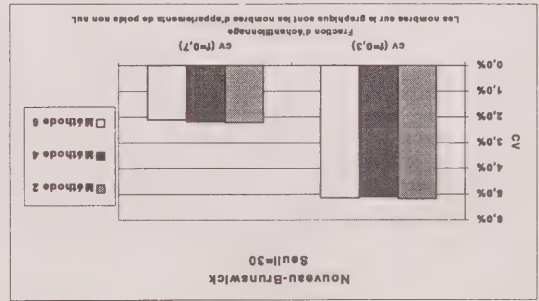
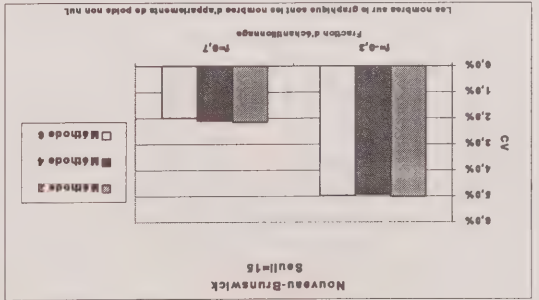
STATISTICS CANADA (2000). *Whole Farm Database reference manual*. Publication No. 21F0005GIE, Statistique Canada, 100 pages.

THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.

THOMPSON, S.K., et SEBER, G.A. (1996). *Adaptive Sampling*. New York: John Wiley and Sons.

WINKLER, W.E. (1995). Matching and record linkage. Dans *Business Survey Methods*, (Éd. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge et P.S. Korn), New York: John Wiley and Sons, 355-384.

YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, 15, 235-261.



Certaines simulations ont été réalisées en se servant des données du Registre des fermes de 1996 (population U^A) et du Fichier des déclarations de revenus des entreprises non constituées en société de 1996 de l'ADRC (population U^B). Nous comparons les variances obtenues pour chaque population par les cinq méthodes : 1) utilisation de tous les appariements de poids non nul; 2) utilisation de tous les appariements de poids non nul et supérieur à un seuil donné; 3) choix aléatoire des appariements d'après des essais de Bernoulli; 4) méthode classique; 5) utilisation de tous les appariements de poids non nul, mais en se servant de la variable indicatrice I . Tous les résultats montrent que les méthodes 1 et 5 sont celles qui produisent les coefficients de variation les plus faibles pour l'estimation du revenu agricole total. Par contre, ces deux méthodes sont celles qui nécessitent le nombre le plus élevé d'appariements ainsi que le nombre le plus élevé de grappes répétées au moyen de s^A , ce qui sous-entend que les coûts de collecte des données sont élevés. Par conséquent, les méthodes 2, 3 et 4 pourraient être considérées comme de bons compromis.

Pour un seuil donné θ^{High} , il est préférable d'utiliser la variable indicatrice I plutôt que le poids de couplage θ pour produire les poids d'estimation selon la MGPP. Ce résultat est confirmé même si $\theta^{\text{High}} = 0$ (c'est-à-dire si l'on n'utilise aucun seuil), comme dans le cas des méthodes 1 et 5. Les estimations produites au moyen de la variable indicatrice I possèdent systématiquement un coefficient de variation plus faible, résultat qui corrobore les conclusions de Kalton et Brick (1995). Donc, la méthode 5 devrait être utilisée de préférence à la méthode 1, et la méthode 4, de préférence à la méthode 2.

L'utilisation du seuil θ^{High} permet de réduire le nombre d'appariements de poids non nul qu'il faut manipuler. En réduisant le nombre d'appariements de poids non nul, nous réduisons aussi le nombre de grappes répétées grâce à l'échantillon s^A , et, donc, les coûts de collecte associés à la mesure de la variable étudiée y dans les grappes. Soulignons que, si nous réduisons le nombre d'appariements, nous diminuons la précision de l'estimation produite. Par conséquent, il faut établir un compromis entre la précision souhaitée et les coûts de collecte des données. La réduction du nombre d'appariements de poids non nul peut aussi être réalisée par application de la règle de décision (2.3) conjuguée aux deux seuils θ^{Low} et θ^{High} . Cette approche réduit les coûts de collecte, mais nécessite certaines interventions manuelles lorsque les poids de couplage θ sont compris entre θ^{Low} et θ^{High} . Les résolutions manuelles de cas produisent néanmoins de meilleurs appariements, c'est-à-dire contenant moins d'erreurs de couplage. L'utilisation de la résolution manuelle pour réaliser les appariements binivoques (un à un) entre les populations U^A et U^B , pourrait ne pas être nécessaire, puisque la MGPP convient particulièrement bien au calcul des estimations dans les situations où les appariements entre

U^A et U^B sont complexes.

Tableau 2
Nombre moyen de grappes répétées

Seuil	Méthode	Nombre moyen de grappes répétées (e.-l.)											
30	Québec	1	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2	11 310(45)	19 139(37)	1 215(17)	1 869(14)	1 924(15)	2 100(7)
		2	11 310(45)	19 139(37)	1 215(17)	1 924(15)	3	10 930(50)	18 881(47)	1 123(14)	1 869(14)	1 924(15)	2 100(7)
		3	10 930(50)	18 881(47)	1 123(14)	1 869(14)	4	11 310(45)	19 139(37)	1 215(17)	1 869(14)	1 924(15)	2 100(7)
		4	14 281(49)	20 593(34)	1 310(17)	1 966(13)	5	15 752(58)	21 106(30)	1 709(18)	1 966(13)	2 100(7)	2 100(7)
		5	15 752(58)	21 106(30)	1 709(18)	2 100(7)	f=3	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2 100(7)	2 100(7)
		f=7	15 752(58)	21 106(30)	1 709(18)	2 100(7)	f=7	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2 100(7)	2 100(7)
15	Nouveau-Brunswick	1	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2	14 281(49)	20 593(34)	1 310(17)	1 966(13)	2 100(7)	2 100(7)
		2	14 281(49)	20 593(34)	1 310(17)	1 966(13)	3	10 930(50)	18 881(47)	1 123(14)	1 869(14)	1 924(15)	2 100(7)
		3	10 930(50)	18 881(47)	1 123(14)	1 869(14)	4	11 310(45)	19 139(37)	1 215(17)	1 869(14)	1 924(15)	2 100(7)
		4	14 281(49)	20 593(34)	1 310(17)	1 966(13)	5	15 752(58)	21 106(30)	1 709(18)	1 966(13)	2 100(7)	2 100(7)
		5	15 752(58)	21 106(30)	1 709(18)	2 100(7)	f=3	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2 100(7)	2 100(7)
		f=7	15 752(58)	21 106(30)	1 709(18)	2 100(7)	f=7	15 752(58)	21 106(30)	1 709(18)	2 100(7)	2 100(7)	2 100(7)



Figure 3. Graphiques des CV en fonction du nombre moyen de grappes répétées

$$(5.5)$$

$$\sum_{M_A} \sum_{N} \sum_{M_B}^{K-1} \sum_{I=1}^{J-1} \tilde{\theta}_{f,ik} = L_0$$

où L_0 est le nombre souhaité d'appariements de poids non nul. Nous avons utilisé la transformation

$$(5.6) \quad \tilde{\theta}_{f,ik} = \begin{cases} \theta_{f,ik}/\theta_{\cdot} & \text{si } \frac{\theta_{f,ik}}{\theta_{\cdot}} \leq 1 \\ 1 & \text{autrement} \end{cases}$$

où θ_{\cdot} a été déterminé de façon itérative de sorte que l'expression (5.5) soit satisfait. Nous baptisons méthode 6 l'utilisation de la méthode 3 conjuguée à la transformation (5.6). Les résultats de la simulation sont présentés aux figures 4.1 à 4.4. Nous constatons que la méthode 6 est celle qui produit les coefficients de variation les plus faibles dans la moitié des cas. Pour les autres, la méthode 4 est s'observe ni pour une province particulière, ni pour une traction d'échantillonnage particulière ni même pour un seuil particulier. Par conséquent, il serait difficile, en pratique, de préciser d'avance si la méthode 6 ou la

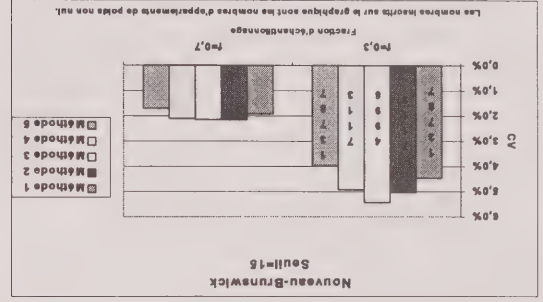


Figure 2.1. CV pour le Nouveau-Brunswick (avec $\theta^{\text{high}} = \theta^{\text{low}} = 15$)

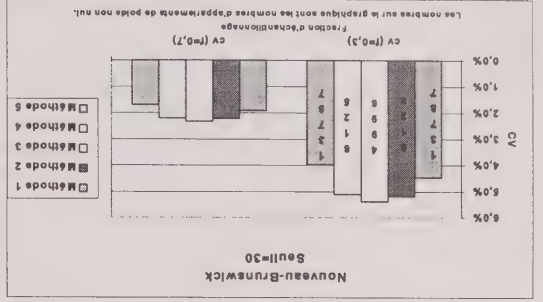


Figure 2.2. CV pour le Nouveau-Brunswick (avec $\theta^{\text{high}} = \theta^{\text{low}} = 30$)

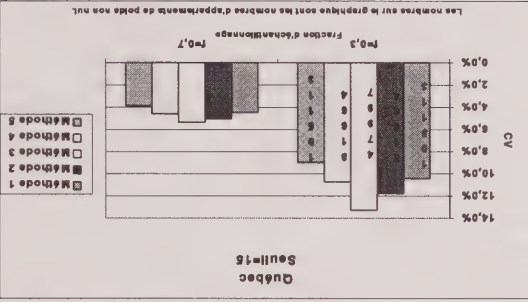


Figure 2.3. CV pour le Québec (avec $\theta^{\text{high}} = \theta^{\text{low}} = 15$)

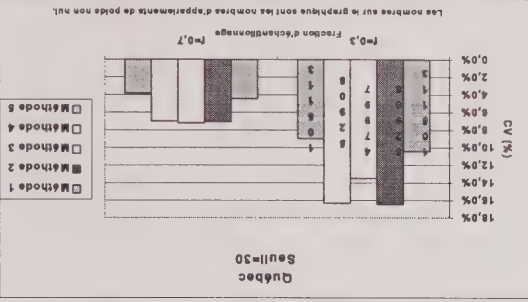


Figure 2.4. CV pour le Québec (avec $\theta^{\text{high}} = \theta^{\text{low}} = 30$)

6. CONCLUSION

Dans le présent article, nous avons montré que la MGPP peut être adaptée à des populations reliées par couplage d'enregistrements. Il s'agit en fait simplement d'une généralisation naturelle du cas où les appariements existent ou n'existent pas, ce qui correspond à l'utilisation d'une variable indicatrice $I_{f,ik} = 1$ si la paire (f, ik) est considérée comme un appariement et 0, autrement. L'union de deux populations par couplage d'enregistrements est toujours entachée d'une certaine incertitude car les décisions concernant les appariements sont fondées sur une méthode probabiliste. Par conséquent, remplacer la variable indicatrice $I_{f,ik}$ par le poids de couplage $\theta_{f,ik}$ qui a été calculé pour chaque paire (f, ik) rend simplement la MGPP plus générale.

méthode 4 produira les coefficients de variation les plus faibles. Pour cette raison et parce que la méthode 6 (et la méthode 3) peut donner lieu à d'importantes erreurs de couplage, on devrait donner la préférence à la méthode 4.

$$CV(\hat{Y}) = 100 \times \frac{\sqrt{V(\hat{Y})}}{\bar{E}(\hat{Y})}. \quad (5.4)$$

Nous avons appliqué la méthode de Monte-Carlo pour vérifier empiriquement l'exactitude des formules théoriques présentées à la section 4. Selon les résultats, toutes les formules théoriques données sont exactes.

Les résultats de l'étude sont présentés aux figures 2.1 à 2.4, au tableau 2 et à la figure 3. Les figures 2.1 à 2.4 donnent des graphiques à barres des coefficients de variation obtenus pour chacune des cinq méthodes. Les graphiques à barres sont produits pour les huit cas obtenus par croisement des deux provinces (Québec et Nouveau-Brunswick), des deux fractions d'échantillonnage (30 % et 70 %) et des deux seuils (15 et 30). Sur chaque barre des graphiques figure le nombre d'appariements de poids non nul entre les unités de U^A et U^B pour chacune des cinq méthodes. Notons que, pour la méthode 3, le nombre correspond en fait au nombre prévu d'appariements de poids non nul. Le nombre (prévu) d'appariements de poids non nul ne varie pas selon la fraction d'échantillonnage. Le tableau 2 montre le nombre moyen de grappes intervenues par méthode, pour chacun des huit cas, lorsque la moyenne est calculée sur les 500 échantillons utilisés pour les simulations. Les chiffres entre parenthèses sont les erreurs-types. Celles-ci sont relativement faibles comparativement aux moyennes et, par conséquent, le nombre de grappes répétées grâce à l'échantillon s^A ne fluctuent pas fortement d'un échantillon s^A à l'autre. La figure 3 donne le diagramme de dispersion des coefficients de variation calculés en fonction du nombre moyen de grappes répétées grâce à l'échantillon s^A , pour chacun des huit cas. Si nous examinons les figures 2.1 à 2.4, nous constatons que, dans tous les cas, la méthode 1 et la méthode 5 sont celles qui produisent les coefficients de variation les plus faibles pour l'estimation du revenu agricole total. Par conséquent, l'utilisation de tous les appariements de poids non nul est la méthode qui donne les résultats les plus précis. Notons cependant que, selon le tableau 2, ces méthodes produisent aussi le nombre le plus élevé de grappes répétées grâce à l'échantillon sélectionné à partir de U^A . En fait, nous constatons que la précision des estimations répétées est d'autant plus grande que le nombre de grappes utilisées pour l'estimation est élevé. Ce résultat est illustré à la figure 3 où nous voyons que les coefficients de variation ont tendance à diminuer à mesure que le nombre moyen de grappes répétées grâce à s^A augmente. Ce résultat est bien connu en théorie classique de l'échantillonnage mais il n'était pas certain qu'il soit vérifié dans le contexte de la MGPP. Comme le montre l'équation (3.5), ce n'est pas la taille de l'échantillon s^A qui augmente, mais plutôt l'homogénéité des variables dérivées Z_j . Maintenant, si nous comparons la méthode 1 et la méthode 5, nous constatons que la seconde produit toujours la variance la plus faible, ce qui donne à penser que l'on devrait utiliser la variable indicatrice 1 plutôt que le poids de

couplage θ lorsque l'on se sert de tous les appariements ayant un poids non nul. Ce phénomène semble généralisé, puisqu'il s'observe aussi si l'on compare la méthode 2 et la méthode 4 (méthode classique). Rappelons que, comme $\theta^{\text{High}} = \theta^{\text{Low}}$, les deux méthodes diffèrent uniquement en ce qui concerne la définition des poids d'estimation obtenus par la MGPP. La méthode 2 se fonde sur les poids de couplage θ , tandis que la méthode classique se fonde sur les variables indicatrices 1. Ces résultats corroborent les conclusions de Kalton et Brick (1995) puisque le choix optimal consistant à permettre à la constante α de prendre la valeur zéro pour certaines unités et une même valeur positive pour toutes les autres unités de la grappe revient à utiliser la variable indicatrice 1. Concentrons-nous maintenant sur la méthode 3. Pour sept des huit histogrammes des figures 2.1 à 2.4, la méthode 3 est celle qui produit les coefficients de variation les plus élevés. Le seul coefficient de variation plus faible est celui obtenu pour le Québec, pour la fraction d'échantillonnage de 30 % et le seuil $\theta^{\text{High}} = 30$. Notons toutefois que cette méthode est celle qui se fonde sur le nombre le plus faible d'appariements ayant un poids non nul, ainsi que sur le nombre moyen le plus faible de grappes répétées grâce à s^A . Par conséquent, ce résultat n'est pas entièrement surprenant. Rappelons que le nombre d'appariements de poids non nul utilisés dans le cas de la méthode 3 ne dépend pas du seuil θ^{High} et que les coefficients de variation obtenus pour le Québec pour $f=0,3$ sont égaux pour $\theta^{\text{High}} = 15$ et $\theta^{\text{High}} = 30$. Pour $\theta^{\text{High}} = 15$, le coefficient de variation obtenu par la méthode 3 pour le Québec est plus élevé que ceux obtenus par les méthodes 2 et 4 qui se fondent sur l'utilisation d'un plus grand nombre d'appariements de poids non nul et d'un plus grand nombre de grappes. Pour $\theta^{\text{High}} = 30$ le coefficient de variation obtenu par la méthode 3 est plus faible que ceux produits par les méthodes 2 et 4, mais ces deux d'appariements de poids non nul et un plus grand nombre de grappes. Il existe donc des situations intermédiaires où, si $15 < \theta^{\text{High}} < 30$, nous devrions obtenir des coefficients de variation égaux pour les méthodes 3 et 2, et pour les méthodes 3 et 4. Par conséquent, pour que les coefficients de variation obtenus par la méthode 3 d'une part, et par la méthode 2 ou 4 d'autre part, soient égaux, on doit utiliser un plus grand nombre d'appariements de poids non nul et de grappes dans le cas de la seconde. Il semble donc que, dans certains cas, la méthode 3 convienne mieux que les méthodes 2 et 4, puisqu'elle permet d'obtenir des estimations de même précision à un coût de collecte moindre. Afin de mieux comparer la méthode 3 aux méthodes 2 et 4, nous avons imposé la condition voulant que le nombre d'appariements de poids non nul attendus soit égal au nombre d'appariements de poids non nul utilisés dans le cadre des méthodes 2 et 4. À cette fin, nous avons transformé les poids de couplage $\theta_{j,k}$ de façon à obtenir $\theta_{j,k}$ afin d'avoir

Tableau 1
Exemple des exploitations agricoles

Nouveau- Brunswick	Taille du Registre des fermes (U^A)	43 017	4 930
	Taille du Fichier des déclarations de revenus (U^B)	52 394	5 155
Nombre total de ménages de U^B			
	22 387	2 194	
Nombre total de poids de couplage			
	105 113	13 787	
non nuls			

Le processus de couplage utilisé pour les simulations consiste en un appariement fondé sur cinq variables. Il a été exécuté en se servant de la commande MERGE de SAS®. Les enregistrements des deux fichiers ont été comparés les uns aux autres afin de déterminer si un appariement éventuel avait eu lieu. Le couplage d'enregistrements a été exécuté en s'appuyant sur les cinq variables clés suivantes, communes aux deux sources :

- prénom (modifié d'après le NYSIS);
- nom de famille (modifié d'après le NYSIS);
- date de naissance;
- adresse de voirie;
- code postal.

Le prénom et le nom de famille ont été modifiés à l'aide du système NYSIS qui, essentiellement, transforme le nom en expressions phonétiques, qui à leur tour augmentent la probabilité de trouver des appariements en réduisant la erreur d'épellation. Pour plus de renseignements sur le NYSIS, consulter Lynch et Arends (1977).

Le poids de couplage le plus élevé ($\theta=60$) a été attribué aux enregistrements appariés sur les cinq variables. Le poids de couplage le plus faible (aussi faible que $\theta=2$) a été attribué aux enregistrements pour lesquels l'appariement ne correspondait qu'à un sous-ensemble d'au moins deux des cinq variables. Il convient de souligner que le niveau des poids de couplage a été choisi arbitrairement. Comme nous l'avons mentionné plus haut, c'est la grandeur relative de ces poids, plutôt que le niveau proprement dit, qui importe. Les paires d'enregistrements ne présentant d'appariement pour aucune combinaison de variables clés n'ont pas été considérées comme des appariements potentiels, ce qui revient à appliquer un poids de couplage nul. Nous avons utilisé deux seuils distincts pour les simulations : $\theta_{High}^{Low} = 15$ et $\theta_{High}^{Low} = 30$. Les seuils supérieur et inférieur, θ_{High}^{Low} et θ_{Low}^{High} , ont été fixés à la même valeur pour éviter la zone grise où certaines interventions manuelles sont nécessaires quand on applique la règle de décision (2.3).

Soulignons qu'il fallait que soit satisfaite la contrainte à laquelle est assujettie l'utilisation de la MGPP. Lorsque, pour une grappe i de U^B , il n'existait aucun poids de couplage $\theta_{j,ik}$ non nul entre toute unité k de cette grappe et les unités de U^A , nous avons imputé un appariement en choisissant celui pour lequel le poids de couplage $\theta_{j,ik}$ était

le plus élevé à l'intérieur de la grappe. Notons en outre que,

pour certaines unités j de U^A , il est arrivé qu'il n'existât aucun poids de couplage $\theta_{j,ik}$ non nul avec toute unité ik de U^B , ce que nous n'avons pas jugé problématique puisque la seule couverture qui nous intéresse est celle de la population U^B . Le tableau 1 donne le nombre total d'appariements de poids non nul observés pour chacune des deux provinces.

Pour les simulations, nous avons sélectionné l'échantillon tiré de U^A (c'est-à-dire le Registre des fermes) par la méthode d'échantillonnage aléatoire simple sans remise (EASSR), sans stratification. Nous avons également constitué deux fractions d'échantillonnage : 30 % et 70 %. La variable étudiée Y^B dont il faut estimer la valeur est le revenu agricole total.

Comme nous disposons de données sur la population complète d'exploitation agricole et de tous les dossiers fiscaux, nous avons pu calculer la variance théorique de ces estimations. En outre, nous avons pu estimer cette variance en sélectionnant un grand nombre d'échantillons (c'est-à-dire par une étude de Monte-Carlo), en estimant le paramètre Y^B pour chaque échantillon, puis en calculant la variance de toutes les estimations. Nous avons appliqué les deux méthodes. Pour les simulations, nous avons sélectionné 500 échantillons aléatoires simples pour chaque méthode, pour les deux fractions d'échantillonnage (30 % et 70 %). Nous avons aussi appliqué les deux seuils de poids (15 et 30) pour mieux comprendre les propriétés des estimateurs donnés.

Puisque nous avons supposé que l'échantillon avait été sélectionné par EASSR, nous avons pu simplifier les formules théoriques présentées à la section 4. Par exemple, dans les conditions d'EASSR, la formule de la variance (4.5) se réduit à l'expression suivante :

(5.1)

$$\text{Var}(Y^{\text{RL}}) = M^A \frac{(1-f)}{f} S_{Z^{\text{RL}}}^2$$

où $f = m^A/M^A$ représente la fraction d'échantillonnage, $S_{Z^{\text{RL}}}^2 = 1/M^A - 1 \sum_{j=1}^{J^{\text{RL}}} (Z_j^{\text{RL}} - \bar{Z}^{\text{RL}})^2$, $\bar{Z}^{\text{RL}} = 1/M^A \sum_{j=1}^{J^{\text{RL}}} Z_j^{\text{RL}}$.

L'étude de Monte-Carlo comprenait 500 répétitions. Pour chacune des deux fractions d'échantillonnage (30 % et 70 %), nous avons sélectionné 500 échantillons aléatoires simples i , puis nous avons estimé l'espérance mathématique et la variance pour chacune des cinq méthodes étudiées en nous servant de

(5.2)

$$E(Y) = \frac{1}{500} \sum_{i=1}^{500} Y_i$$

et de

(5.3)

$$V(Y) = \frac{1}{500} \sum_{i=1}^{500} (Y_i - E(Y))^2.$$

Nous avons calculé comme suit les coefficients de variation (CV) estimatifs :

Rappelons que l'exemple des exploitations agricoles est celui d'une enquête auprès de ces exploitations agricoles où la première population U^A est une liste d'exploitations déterminées d'après le Recensement de l'agriculture du Canada. Cette liste provient du Registre des fermes de 1996 qui est essentiellement la liste de tous les enregistrés durant le Recensement de l'agriculture de 1991 et de toutes les mises à jour qui ont eu lieu depuis. Il contient un identificateur d'exploitants agricoles, ainsi que certaines variables sociodémographiques en rapport avec les exploitants agricoles. La deuxième population U^B est une liste de dossiers fiscaux produite par l'IDRC. Cette deuxième liste est le Fichier des déclarations de revenus des entreprises non constituées en société de 1996 produit par l'ADRC qui contient les données sur les personnes qui déclarent au moins un revenu agricole. Il contient un identificateur de ménage (uniquement pour un échantillon) un identificateur de déclarant, ainsi que des variables sociodémographiques en rapport avec les déclarants.

À Statistique Canada, la Division de l'agriculture produit des estimations des récoltes et du bétail d'après des échantillons tirés du Registre des fermes (population U^A). Pour créer la Base de données complètes sur les exploitations agricoles, il faut recueillir des données fiscales sur les exploitations agricoles qui ont été sélectionnées dans les échantillons du Registre des fermes. On commence, pour cela, par fusionner le Registre des fermes et le Fichier des déclarations de revenus des entreprises non constituées en société (population U^B), puis on obtient les données fiscales provenant de l'ADRC. Comme nous l'avons mentionné plus haut, la relation entre les exploitants agricoles enregistrés dans le Registre des fermes et les déclarants enregistrés dans le Fichier des déclarations de revenus des entreprises non constituées en société de l'ADRC n'est pas biunivoque (un à un). C'est pour cette raison que la MGPP est une bonne méthode pour produire les poids d'estimation à appliquer aux déclarants sélectionnés d'après l'échantillon d'exploitants agricoles provenant du Registre des fermes.

D'aucuns soutiendront qu'il n'est pas nécessaire d'obtenir un ensemble de grappes repérées au moyen des unités $f \in s^A$, puisque la population cible U^B est une population de déclarants extraite du Fichier des déclarations de revenus des sociétés non constituées en société de l'ADRC qui est habituellement produit par le recensement. Notons

toutefois que cet argument n'est pas entièrement correct. Les variables étudiées ne figurent pas toutes dans ce fichier et Statistique Canada doit payer pour les renseignements sur les variables supplémentaires que lui fournit l'ADRC. En outre, les données provenant du Fichier des déclarations de revenus des entreprises non constituées en société contiennent des erreurs de saisie clavier, de codage, etc. qui entraînent certains frais d'épuration des données. Il est donc préférable de limiter les données provenant de la population cible U^B à un sous-ensemble uniquement. Or, un moyen de déterminer l'ensemble de grappes qu'il faut utiliser pour estimer Y^B consiste simplement à le faire d'après l'échantillon s^A sélectionné à partir de U^A .

Sauf dans le cas de la méthode classique, le couplage proprement dit entre U^A et U^B est considéré pour toutes les méthodes étudiées comme un objectif secondaire. L'objectif principal étant de produire une estimation Y^B pour la population cible U^B . Cependant, l'application mentionnée ici est reliée à la Base de données complètes sur les exploitations agricoles conçue pour être une base de données intégrées. Si le couplage entre les enregistrements sur les populations U^A et U^B n'est pas de bonne qualité, les analyses des microdonnées sur les variables de récolte et de détail obtenues d'après l'échantillon s^A et d'après le fichier de données fiscales provenant de U^B seront erronées. À cet égard, nous reconnaissons que les méthodes proposées, sauf la méthode classique, ne sont pas viables dans le contexte actuel. Elles le sont cependant si on les examine dans une perspective à long terme. Comme des interventions manuelles sont nécessaires lorsque l'on utilise une règle de décision telle que (2.3), on pourrait suggérer d'utiliser les méthodes proposées pour produire d'après U^B , à court terme, certaines estimations requises, avant que les résultats finals du couplage soient disponibles, après la résolution manuelle de certains cas. Rappelons que l'objectif principal de la simulation est d'évaluer les méthodes proposées en regard de la méthode classique. L'exemple des exploitations agricoles n'a pas été choisi parce qu'il correspond à une situation réelle, mais parce que les données existent. On aurait pu prendre n'importe quel autre exemple, comme celui également mentionné dans l'introduction où U^A est une population de parents et U^B , une population d'enfants appartenant à ces parents.

Aux fins des simulations, nous avons choisi deux provinces du Canada, à savoir le Nouveau-Brunswick que nous considérons comme une petite province, et le Québec, qui est une grande province. Le tableau 1 donne la taille des différents fichiers. Comme l'identificateur de ménage n'existe pas pour la population U^B complète, aux fins des simulations, nous l'avons créé pour un échantillon pour lequel le numéro d'identification de ménage a été codé pour chaque déclarant. Pour les déclarants non échantillonnés, l'identificateur de ménage a été attribué aléatoirement de sorte que les proportions des diverses tailles de ménages correspondent à celles observées pour l'échantillon.

4.4 Quelques remarques

différents : celui obtenu par la méthode classique dépend des seuils $\theta_{f,ik}^{Low}$ et $\theta_{f,ik}^{High}$, tandis que celui obtenu par la méthode 3 dépend des poids de couplage corrigés $\theta_{f,ik}$ qui correspondent aux probabilité de sélection des appariements.

5. ÉTUDE EN SIMULATION

Nous avons exécuté une étude en simulation pour évaluer les méthodes proposées comparativement à la méthode classique fondée sur l'application de la règle de décision (2.3) pour déterminer quels sont les appariements vrais. Pour réaliser l'étude, nous avons comparé la précision de l'estimation d'un total Y^B obtenu en appliquant les cinq méthodes suivantes :

- Méthode 1** : utilisation de tous les appariements de poids de couplage non nul en appliquant les poids de couplage respectifs
- Méthode 2** : utilisation de tous les appariements de poids non nul et supérieur à un seuil donné
- Méthode 3** : choix aléatoire des appariements d'après des essais de Bernoulli
- Méthode 4** : méthode classique
- Méthode 5** : utilisation de tous les appariements de poids non nul, mais en servant de la variable indicatrice I .

La méthode 5 est une combinaison de la méthode 1 et de la méthode classique. Elle consiste fondamentalement à accepter comme étant des appariements vrais toutes les paires (j, ik) dont le poids de couplage n'est pas nul, c'est-à-dire à supposer que $I_{j,ik} = 1$ pour toutes les paires (j, ik) où $\theta_{f,ik} > 0$, et 0 autrement, puis à utiliser la MGPP décrite à la section 3 pour produire l'estimation de Y^B . La méthode 5 a été ajoutée à l'étude en simulation pour observer l'effet de la substitution de la variable indicatrice I au poids de couplage θ lorsque l'on utilise tous les appariements de poids non nul. Comme pour les autres méthodes, on peut montrer que la méthode 5 est non biaisée.

Étant donné que les cinq méthodes produisent des estimations du total Y^B , sans biais dû au plan d'échantillonnage, nous avons choisi comme paramètre de comparaison l'erreur-type de l'estimation ou simplement le coefficient de variation (c'est-à-dire le ratio de la racine carrée de la variance à la valeur prévue). L'étude en simulation se fonde sur l'exemple des exploitations agricoles mentionné tout au long de l'article. Cet exemple correspond en fait à une situation réelle vécue à Statistique Canada lors de la création de la Base de données complètes sur les exploitations agricoles (voir Statistique Canada 2000). Bien que l'étude en simulation se fonde sur la situation réelle, certains chiffres ont été modifiés pour des raisons de confidentialité. En outre, la méthode de

Les trois méthodes proposées ne s'appuient aucune sur la règle de décision (2.3). En outre, elles ne nécessitent pas de résolution manuelle. Donc, la réponse à la question (c) permet de réduire le nombre d'interventions manuelles que demande le couplage d'enregistrements. Notons, cependant, qu'il n'est possible d'éviter la résolution manuelle qu'à un certain prix.

En premier lieu, si l'on utilise la méthode 1, le nombre n_{RL} de grappes repérées sur les unités $j \in \mathcal{A}$ est égal ou supérieur au nombre n de grappe repérées par la méthode classique, c'est-à-dire lorsqu'on applique la règle de décision (2.3) pour repérer les appariements. Cette situation tient au fait que nous utilisons tous les appariements de poids non nul plutôt que simplement ceux qui répondent aux critères de la règle de décision (2.3). Par conséquent, le recours à la méthode 1 occasionnera des frais de collecte de données égaux ou supérieurs à ceux associés à la méthode classique. Il convient donc de déterminer si ce sont les coûts de collecte de données qui sont les plus élevés. Notons que, si la méthode 1 produit des résultats nettement plus précis que la méthode classique, il pourrait être plus intéressant d'utiliser la première que la seconde.

Si nous appliquons la méthode 2, nous obtenons $n_{RL} \leq n_{RL}$ et par conséquent, les coûts de collecte sont égaux ou inférieurs à ceux associés à la méthode 1. Si la précision de la méthode 2 est comparable à celle de la méthode 1, alors la première sera manifestement plus avantageuse que la seconde. Si nous comparons la méthode 2 à la méthode classique, nous constatons que les coûts de collecte des données peuvent être presque équivalents si la valeur du seuil θ_{High} est choisie de façon à ce qu'elle soit proche des seuils inférieur et supérieur de la règle de décision (2.3). Notons que la méthode 2 ne demande aucune résolution manuelle de cas. Si sa précision est au moins comparable à celle de la méthode classique, alors elle présente un avantage net. Notons aussi que, si $\theta_{High} = \theta_{Low}$, les deux méthodes ne diffèrent qu'en ce qui concerne la définition des poids d'estimation obtenus par la MGPP. La méthode 2 se fonde sur le poids de couplage θ , tandis que la méthode classique se fonde sur la variable indicatrice I . Quand on pose que $\theta_{High} = \theta_{Low}$, il est sans aucun doute intéressant de déterminer quelle méthode est la plus précise. Dans le cas de la méthode 3, le nombre d'appariements sélectionnés sera inférieur ou égal au nombre d'appariements de poids non nul utilisés lors de l'application de la méthode 1, c'est-à-dire $n \leq n_{RL}$. Donc, les coûts de collecte de données de la méthode 3 seront inférieurs ou égaux à ceux de la méthode 1. Quant à la précision, il n'est pas possible d'établir clairement laquelle des deux méthodes est susceptible de donner la variance la plus faible. Comme nous l'avons mentionné plus haut, contrairement à n_{RL} et n_{RL} le nombre aléatoire de grappes n n'est guère comparable à n . Ces deux nombres dépendent de paramètres

correspondent en fait à une transformation logit (en base 2) de la probabilité $P(\mu^k | C_{1k} C_{2k} \dots C_{Qk})$. Pareillement, les poids de couplage donnés par l'équation (2.2) dépendent uniquement de cette probabilité. Par conséquent, un moyen de transformer les poids de couplage consiste simplement à utiliser la probabilité $P(\mu^k | C_{1k} C_{2k} \dots C_{Qk})$. D'après (2.1), nous obtenons ces résultats en utilisant la fonction $\theta = 2^q / (1 + 2^q)$. D'après (2.2), nous utilisons $\theta = \theta / (1 + \theta)$. Si les poids de couplage ne sont obtenus au moyen ni de (2.1) ni de (2.2), une transformation possible consiste à diviser chaque poids de couplage par la valeur maximale possible $\theta_{\max} = \max_{j=1, \dots, i-1, k=1, \dots, M_j^A} \theta_{j,ik}$. Notons que nous supposons que les poids de couplage ont tous une valeur égale ou supérieure à zéro, ce qui est le cas pour la définition (2.2), mais pas nécessairement de façon générale.

Une fois que nous avons calculé les poids de couplage corrigés $\tilde{\theta}_{j,ik}$, nous produisons pour chaque paire (j, ik) , un nombre aléatoire $u_{j,ik} \sim U(0,1)$. Puis, nous fixons la valeur de la variable indicatrice $I_{j,ik}$ à 1 si $u_{j,ik} \leq \tilde{\theta}_{j,ik}$, et à 0, autrement. Cette méthode produit un ensemble d'appariements similaires à ceux utilisés dans la méthode classique, à part qu'ici, les appariements sont déterminés de façon aléatoire plutôt que par application d'une règle de décision comparable à (2.3). Notons que, puisque $E(I_{j,ik}) = \theta_{j,ik}$, la somme des poids de couplage corrigés $\tilde{\theta}_{j,ik}$ est égale au nombre total prévu L d'appariements dans $A \times B$, calculé par la méthode de Bernoulli, c'est-à-dire

$$(4.8) \quad \sum_{M^A} \sum_{N} \sum_{M^B} \tilde{\theta}_{j,ik} = L.$$

Pour chaque unité j sélectionnée dans s^A , nous repérons les unités ik de U^B pour lesquelles $I_{j,ik} = 1$. Représentons par Ω^B l'ensemble des \tilde{n} grappes repérées pour les unités $j \in s^A$. Notons que $\tilde{n} \leq n_{RL}$. Malheureusement, contrairement à n_{RL} et n_{RLT} , le nombre aléatoire de grappes \tilde{n} n'est guère comparable à n .

Nous définissons le poids initial \tilde{w}_i^k comme suit :

$$(4.9) \quad Y = \sum_{M^A} \sum_{N} \sum_{M^B} \frac{\pi_j^A}{t_j} \sum_{k=1}^{\tilde{n}} I_{j,ik} z_{ik} = \sum_{M^A} \sum_{t_j} \frac{\pi_j^A}{t_j} Z_j.$$

Le poids final \tilde{w}_i^k est donné par

$$(4.10) \quad \tilde{w}_i^k = \frac{\sum_{M^B} \sum_{t_j} \frac{\pi_j^A}{t_j} I_{j,ik}}{\sum_{M^B} \sum_{k=1}^{\tilde{n}} I_{j,ik}}.$$

La quantité $\tilde{L}_{j,ik}$ représente le nombre réalisé d'appariements entre les unités de U^A et l'unité k de la grappe i de la population U^B . Enfin, nous supposons que $\tilde{w}_{ik}^i = w_i^i$ pour toutes les unités $k \in U_i^i$.

Pour estimer le total Y^B , nous pouvons utiliser l'estimateur

$$(4.11) \quad \hat{Y} = \sum_{M^B} \sum_{k=1}^{\tilde{n}} \sum_{i=1}^I \tilde{w}_{ik}^i Y_{ik}^i.$$

En établissant les conditions en fonction des appariements acceptés $I_{j,ik}$, nous pouvons montrer que l'estimateur (4.11) est conditionnellement dépourvu de biais dû au plan d'échantillonnage et, donc, est inconditionnellement dépourvu d'un biais de ce genre. Notons que, comme les conditions introduites sont axées sur $I_{j,ik}$, cet estimateur est équivalent à (3.1). Pour obtenir la variance de \hat{Y} , il faut de nouveau recourir à des arguments conditionnels. Si nous utilisons l'indice I pour indiquer que l'espérance est calculée sur tous les ensembles possibles d'appariements, nous obtenons

$$(4.12) \quad \text{Var}(\hat{Y}) = E_1 \text{Var}_2(\hat{Y}) + \text{Var}_1 E_2(\hat{Y}).$$

En premier lieu, en vertu de l'absence conditionnelle de biais, nous obtenons

$$(4.13) \quad E_2(\hat{Y}) = Y^B.$$

Par conséquent,

$$(4.14) \quad \text{Var}_1 E_2(\hat{Y}) = 0.$$

En deuxième lieu, partant de (3.5), nous obtenons

$$(4.15) \quad \text{Var}_2(\hat{Y}) = \sum_{M^A} \sum_{M^B} \sum_{j=1}^I \sum_{k=1}^{\tilde{n}} \frac{(\pi_j^A - \pi_j^A \pi_j^B)}{(\pi_j^A - \pi_j^A \pi_j^B)} Z_j Z_j^B.$$

où Z_j est défini comme dans (3.4), mais en remplaçant les appariements I par I . Par conséquent, la variance de \hat{Y} peut être exprimée par

$$(4.16) \quad \text{Var}_2(\hat{Y}) = E_1 \left(\sum_{M^A} \sum_{M^B} \sum_{j=1}^I \sum_{k=1}^{\tilde{n}} \frac{\pi_j^A \pi_j^B}{(\pi_j^A - \pi_j^A \pi_j^B)} Z_j Z_j^B \right)$$

où l'espérance est calculée sur tous les ensembles possibles d'appariements.

Dans le cas de la MGPP, nous avons énoncé à la section 3 une contrainte qui doit être satisfaite pour que la méthode ne soit pas biaisée. Dans le cas de l'approche exposée ici, il est fort probable, comme nous sélectionnons aléatoirement les appariements, que cette contrainte ne soit pas satisfaite. Pour surmonter ce problème, nous pouvons imputer un appariement en choisissant celui qui, dans la grappe, possède le poids de couplage $\theta_{j,ik}$ non nul le plus élevé. Si nous n'obtenons toujours aucun appariement parce que tous les $\theta_{j,ik} = 0$, il est possible de choisir un appariement au hasard dans la grappe. Il convient de souligner que, si l'on recourt à cette solution, la MGPP demeure dépourvue de biais dû au plan d'échantillonnage.

l, pourrait nécessiter la manipulation de grands fichiers de taille $M^A \times M^B$, parce que la plupart des enregistrements des fichiers A et B pourraient avoir un poids de couplage non nul. En pratique, si la situation se présente, nous pouvons nous attendre à ce que les poids de couplage soient assez faibles, voire négligeables, de sorte que, même s'ils ne sont pas nuls, il est fort probable que les appartements observés ne soient pas des appartements vrais. Il pourrait alors être utile de ne considérer que les appartements dont le poids de couplage θ est supérieur à un seuil θ_{High} .

Une fois de plus, pour cette deuxième méthode, nous ne nous servons plus de la variable indicatrice l_{jk} précisant s'il y a ou non appartenance vrai, mais plutôt des poids de couplage θ_{jk} dont la valeur est supérieure au seuil θ_{High} . Les poids de couplage dont la valeur est inférieure au seuil sont considérés comme étant nuls. Par conséquent, nous définissons le poids de couplage comme suit :

$$\theta_{jk} = \begin{cases} \theta_{jk} & \text{si } \theta_{jk} \geq \theta_{High} \\ 0 & \text{autrement.} \end{cases}$$

Pour chaque unité j sélectionnée dans s^A , nous repérons les unités ik de U^B pour lesquelles $\theta_{jk}^* > 0$. Représentons par $\Omega_{RLT,B}$ l'ensemble des n_{RLT} grappes repérées pour les unités $j \in s^A$, ou « RLT » signifie « Record Linkage with Threshold », c'est-à-dire couplage d'enregistrements avec seuil. Notons que $n_{RLT} \leq n_{RL}$. Par ailleurs, nous avons $n_{RLT} = n$ si nous réalisons le couplage d'enregistrement entre U^A et U^B en appliquant la règle de décision (2.3) et en posant que $\theta_{High} = \theta_{Low}^*$.

Le poids initial w_{RLT}^* est donné par

$$w_{RL}^{*ik} = \sum_{j=1}^{M^A} \theta_{jk}^* \frac{\pi_j^A}{f_j}. \quad (4.6)$$

Le poids final w_{RLT}^* est donné par

$$w_{RL}^* = \frac{\sum_{k=1}^{M^B} \theta_{jk}^*}{\sum_{k=1}^{M^B} w_{RLT}^*}. \quad (4.7)$$

ou $\Theta_{jk}^* = \sum_{j=1}^{M^A} \theta_{jk}^*$. Enfin, nous posons que $w_{RLT}^* = w_{RLT}^*$ pour toutes les unités $k \in U_{jk}^B$. Comme dans le cas de la méthode 1, il est intéressant de souligner que, si nous posons que $\alpha_{jk}^* = \theta_{jk}^* / \Theta_{jk}^*$ ou $\Theta_{jk}^* = \sum_{j=1}^{M^A} \theta_{jk}^*$, nous obtenons pour le poids d'estimation w_{RLT}^* , une formulation équivalente à celle donnée par (3.7) et (3.8).

Le nombre de poids de couplage θ_{jk}^* dont la valeur est nulle sera supérieur ou égal au nombre de poids de couplage θ nuls utilisés dans le cas de la méthode 1. Par conséquent, la condition voulant qu'il existe pour chaque grappe i de U^B au moins un poids de couplage θ_{jk}^* avec une unité j de U^A de valeur non nulle pourrait être plus difficile à satisfaire. Le cas échéant, l'utilisation du poids

d'estimation (4.7) produit une sous-estimation du total Y^B . Pour résoudre ce problème, nous pouvons recourir aux mêmes solutions que celles proposées plus haut.

Pour estimer le total Y^B , nous pouvons utiliser le même estimateur que (4.3), où nous remplaçons le nombre de grappes repérées n_{RL} par n_{RLT} et le poids d'estimation w_{RL}^* par w_{RLT}^* . Comme pour l'estimateur Y_{RLT}^B est dépourvu de biais dû au plan de sondage.

4.3 Méthode 3 : Choisir les appartements par sélection aléatoire

Afin de ne pas devoir décider s'il y a ou non un appartenance vrai entre une unité j de U^A et une unité k de la grappe i de U^B , nous pouvons simplement choisir les appartements au hasard parmi l'ensemble d'appartements de poids non nul. Il est raisonnable, pour cela, de choisir les appartements avec une probabilité proportionnelle au poids de couplage θ . Nous pouvons pour cela procéder à des essais de Bernoulli, où, pour chaque paire (j, ik) , nous décidons s'il s'agit ou non d'un appartenance vrai en générant un nombre aléatoire $u_{jk} \sim U(0,1)$ que nous comparons au poids de couplage θ_{jk} .

Du point de vue du couplage d'enregistrements, nous ne pouvons considérer cette méthode comme étant optimale. L'intention, lorsque l'on applique la règle de décision (2.3) de Fellegi et Sunter, est de réduire au minimum le nombre de faux appartements et de faux non-appartements. Le lien l_{jk} n'est accepté que si le poids de couplage θ_{jk} est suffisamment grand (c'est-à-dire $\theta_{jk} \geq \theta_{High}$) ou bien s'il est modérément grand (c'est-à-dire $\theta_{jk} > \theta_{Low}^*$) et $\theta_{jk} < \theta_{High}$) qu'il a été accepté après un examen manuel. Choisir aléatoirement les appartements par essais de Bernoulli pourrait mener à la sélection d'appartements qui seraient rejetés par application de la règle de décision (2.3), même si les probabilités de sélection sont proportionnelles au poids de couplage. Le cas échéant, certains appartements entre les deux populations U^A et U^B pourraient être faux et certaines unités qui ne sont pas appartées pourraient représenter de faux non-appartements. Par conséquent, les erreurs de couplage seront vraisemblablement plus nombreuses que lorsque l'on applique la règle de décision (2.3). Cependant, ici, la qualité du couplage ne présente qu'un intérêt secondaire. Le problème qui se pose est celui de l'estimation du total Y^B d'après l'évaluation s^A sélectionnée à partir de U^A plutôt que celui de l'évaluation de la qualité des appartements. La précision des estimations de Y^B ne sera, en fait, mesurée qu'en fonction de la variabilité d'échantillonnage des estimateurs, en imposant comme condition les poids de couplage θ_{jk} . Notons que cette variabilité d'échantillonnage tiendra compte de la sélection aléatoire des appartements, mais non des erreurs de couplage.

La première étape, avant d'exécuter les essais de Bernoulli, consiste à transformer les poids de couplage afin de les limiter à l'intervalle $[0,1]$. Si nous examinons l'équation (2.1), nous constatons que les poids de couplage

(j, ik), avec $\alpha_i = \sum_{j=1}^{M_A} \sum_{k=1}^{M_B} \alpha_{j,ik} = 1$. Nous obtenons alors de nouveaux poids d'estimation comme suit. À chaque unité k de la grappe i entrant dans Y , nous attribuons le poids initial w_{ik}^a suivant :

$$(3.7) \quad w_{ik}^a = \sum_{j=1}^{M_A} \alpha_{j,ik} \frac{\pi_{j,A}}{t}.$$

Le poids final w_i^a est donné par

$$(3.8) \quad w_i^a = \sum_{k=1}^{M_B} w_{ik}^a = \sum_{k=1}^{M_B} \sum_{j=1}^{M_A} \alpha_{j,ik} \frac{\pi_{j,A}}{t}.$$

Enfin, nous supposons que $w_a^k = w_i^a$ pour toutes les unités $k \in U_B^i$ et utilisons l'équation (3.1) pour estimer le total Y_B^i .

Dans le contexte des enquêtes longitudinales, Ernst

(1989) a fait remarquer que la valeur choisie le plus souvent pour les constantes α est celle pour laquelle chaque unité se voit attribuer l'une de deux valeurs, à savoir 0 ou une valeur non nulle qui est la même pour toutes les unités qui restent dans la grappe. Ici, cela reviendrait à supposer que $\alpha_{j,ik} = 0$ pour toutes les unités j et k dans un sous-ensemble U_{0B}^i de U_B^i , disons, et $\alpha_{j,ik} = 1$ pour toutes les unités j et k comprises dans le sous-ensemble complémentaire U_{0B}^i . De nouveau

Brick (1995) se sont penchés sur la détermination des valeurs optimales de α de Ernst (1989). L'optimalité étant déterminée par la variance minimale. Ils concluent que, dans le cas de deux ménages, le scénario de pondération égale des ménages minimise la variance des poids appliqués aux ménages autour du poids correspondant à la probabilité inverse de sélection lorsque l'échantillon initial est sélectionné par mepc. (Méthode d'échantillonnage avec probabilités égales.) Ils ajoutent aussi que, dans le cas d'un échantillon approuvativement mepc, le scénario de pondération égale des ménages devrait être proche du scénario optimal, au moins dans le cas où les membres du ménage observés à la période t proviennent d'un ou de deux ménages sélectionnés lors du cycle initial. Ceci donne à penser que, pour la MGPP, choisir de permettre aux constantes α d'avoir la valeur 0 pour certaines unités et une même valeur positive pour toutes les autres unités de la

grappe devrait s'approcher du scénario optimal.

4. LA MGPP ET LE COUPLAGE D'ENREGISTREMENTS

Par couplage d'enregistrements, on établit les liens, ou appartenements, $l_{j,ik}$ entre les fichiers A et B, ou entre les populations U_A^j et U_B^k , selon une méthode probabiliste. Comme nous l'avons mentionné plus haut, le couplage d'enregistrements se fonde sur l'utilisation d'une règle de décision D telle que (2.3) pour déterminer s'il y a ou non appartenement entre l'unité j du fichier A et l'unité ik du

plusieurs à plusieurs).

Même en cas d'appariements complexes, nous pouvons appliquer la MGPP pour estimer le total Y_B^i pour la population U_B^i en nous servant de l'échantillon s^A tiré de la population U_A^j . Par conséquent, la réponse à la question (a)

énoncée dans l'introduction est affirmative. Notons toutefois que les estimations produites par application de la MGPP pourraient ne pas être dépourvues de biais si la contrainte mentionnée à la section 3 n'est pas satisfaite. Le cas échéant, l'utilisation du poids d'estimation (3.3) produit une sous-estimation du total Y_B^i . Pour résoudre ce problème, une solution pratique consiste à regrouper deux grappes afin d'obtenir au moins un appariement à valeur non nulle $l_{j,ik}$ pour la grappe i . Cette solution demande en général une certaine intervention manuelle. Une autre solution consiste à imputer un appariement en en choisissant un au hasard dans la grappe ou en choisissant celui caractérisé par le poids de couplage $\theta_{j,ik}$ le plus élevé. Il se pourrait cependant qu'il n'existe aucun lien $l_{j,ik}$ non nul entre une unité j de U_A^j , et toute unité ik de U_B^i . Une telle situation ne serait pas problématique, puisque la seule couverture qui nous intéresse ici est celle de U_B^i .

Il est maintenant clair que l'on peut utiliser la MGPP dans le contexte du couplage d'enregistrements. Dans la suite de l'article, nous appellerons méthode classique l'application de la MGPP aux populations U_A^j et U_B^i unies par couplage d'enregistrements conformément à la règle de décision (2.3).

Dans le cas de la méthode classique, l'utilisation de la MGPP se fonde sur des appariements repérés au moyen de la variable indicatrice $l_{j,ik}$. Or, est-il nécessaire de déterminer s'il existe ou non un appariement positif pour chaque paire (j, ik)? Ne serait-il pas plus facile de se servir simplement des poids de couplage $\theta_{j,ik}$ (sans appliquer la règle de décision) pour estimer le total Y_B^i pour U_B^i d'après un échantillon tiré de U_A^j ? Ces questions mènent à la question (b) de l'introduction, c'est-à-dire celle de savoir s'il est possible ou non d'adapter la MGPP pour tenir compte des poids de couplage θ dérivés du couplage d'enregistrements.

À la présente section, nous montrons que la réponse à la question (b) est affirmative en examinant trois méthodes où l'application de la MGPP se fonde sur les poids de couplage θ . La première consiste à utiliser tous les appariements de poids non nul repérés par couplage d'enregistrements en leur appliquant leur poids de couplage θ respectif. La deuxième est celle où l'on utilise tous les appariements à poids de couplage non nul dont la valeur est supérieure à un seuil θ_{High} donné. La troisième est celle où les appariements sont choisis au hasard avec probabilité proportionnelle au poids de couplage θ .

des M_i^t unités de la grappe i contenant cette unité. Alors, chaque grappe i représente, en soi, une population U_i^t où $U_B = \bigcup_{i=1}^N U_i^t$. Représentons par Ω^B l'ensemble des n grappes repérées pour les unités $j \in s^A$.

Pour la population U_B , nous cherchons à estimer le total $Y_B = \sum_{i=1}^N \sum_{k=1}^{M_i^B} y_{ik}$ pour une caractéristique y . Une contrainte importante à laquelle nous assujettissons la méthode de mesure (ou d'interview) de y consiste à tenir compte de toutes les unités appartenant à une même grappe. Autrement dit, si une unité est sélectionnée dans l'échantillon, alors toutes les unités de la grappe contenant cette unité seront interviewées. Cette contrainte est souvent imposée dans les enquêtes par souci d'économie et par nécessité de produire des estimations sur des grappes. Par exemple, dans le cas des enquêtes sociales, le fait d'interviewer tous les membres du ménage entraîne habituellement un léger coût marginal. Toutefois, les estimations produites au niveau du ménage présentent souvent un intérêt pour ceux qui cherchent à évaluer la pauvreté, par exemple. Dans l'exemple des exploitations agricoles, l'une des valeurs qui présente un intérêt est le revenu agricole total par ménage. Pour pouvoir le calculer, nous devons interviewer tous les membres du ménage.

En appliquant la MGPP, nous cherchons à attribuer un poids d'estimation w_{ik} à chaque unité k d'une grappe interviewée i . Pour estimer le total Y_B^t pour la population U_B , nous pouvons alors nous servir de l'estimateur

$$\hat{Y}^t = \sum_{i=1}^n \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \quad (3.1)$$

où n est le nombre de grappes interviewées et w_{ik} le poids attribué à l'unité k de la grappe i . Dans le cas de la MGPP, nos méthodes d'estimation du total Y_B^t se fondent sur l'échantillon s^A , ainsi que sur les liens qui existent entre U^A et U_B . En fait, ces liens forment un pont qui permet de passer de la population U^A à la population U_B , et inversement.

La MGPP vise à attribuer à chaque unité interviewée ik un poids final correspondant à la moyenne des poids calculés à l'intérieur de chaque grappe i entrant dans Y . On calcule d'abord un poids initial, qui correspond à l'inverse de la probabilité de sélection, pour toutes les unités k de la grappe i de Y ayant un lien non nul avec une unité $j \in s^A$. On attribue un poids initial nul aux unités pour lesquelles il n'existe aucun lien. On obtient le poids final en calculant le ratio de la somme des poids initiaux pour la grappe sur le nombre total de liens, ou séparément, pour cette grappe. Enfin, on applique ce poids final à toutes les unités de la grappe. Notons qu'attribuer le même poids d'estimation à toutes les unités à l'important avantage d'assurer la cohérence des estimations calculées pour les unités et pour les grappes.

Formellement, nous attribuons à chaque unité k de la grappe i entrant dans Y un poids initial w_{ik}^* , à savoir :

$$w_{ik}^* = \sum_{j=1}^{M_i^A} l_{j,ik} \frac{\pi_j^A}{t_j} \quad (3.2)$$

où $t_j = 1$ si $j \in s^A$ et 0, autrement. Soulignons qu'une unité ik n'ayant de lien avec aucune unité j de U^A possède automatiquement un poids initial nul. Le poids final w_i^t est donné par

$$w_i^t = \frac{\sum_{k=1}^{M_i^B} w_{ik}^*}{\sum_{k=1}^{M_i^B} L_{ik}} \quad (3.3)$$

où $L_{ik} = \sum_{j=1}^{M_i^A} l_{j,ik}$. La quantité L_{ik} représente le nombre d'appariements entre les unités de U^A et l'unité k de la grappe i de la population U_B . La quantité $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ correspond alors au nombre total d'appariements dans la grappe i . Enfin, nous supposons que $w_{ik} = w_i^t$ pour toutes les unités $k \in U_B^t$ et nous nous servons de l'équation (3.1) pour estimer le total Y_B^t .

En partant de cette dernière expression, Lavallée (1995) a montré que la MGPP ne présente aucun biais dû au plan de sondage. Supposons, en outre, que $z_{ik} = Y_i/L_i$ pour toutes les unités $k \in i$, où $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$. Alors, nous pouvons exprimer Y sous la forme

$$Y = \sum_{i=1}^{M_A} \sum_{j=1}^N \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M_A} \sum_{i=1}^N \frac{\pi_j^A}{t_j} Z_j \quad (3.4)$$

et la variance de Y sous la forme

$$\text{Var}(Y) = \sum_{j=1}^{M_A} \sum_{i=1}^N \frac{\pi_j^A \pi_{ij}^A}{(\pi_{ij}^A - \pi_j^A \pi_{ij}^A)} Z_j Z_{ij} \quad (3.5)$$

où π_{ij}^A représente la probabilité conjointe de sélectionner les unités j et i . Consulter Särndal, Swensson et Wretman (1992) pour le calcul de π_{ij}^A dans le cas de divers plans d'échantillonnage. Nous pouvons estimer sans biais la variance $\text{Var}(Y)$ d'après l'équation :

$$\text{Var}(Y) = \sum_{j=1}^{M_A} \sum_{i=1}^N \frac{\pi_j^A \pi_{ij}^A}{(\pi_{ij}^A - \pi_j^A \pi_{ij}^A)} t_j Z_j t_{ij} Z_{ij}. \quad (3.6)$$

Un autre estimateur non biaisé de la variance $\text{Var}(Y)$ peut être développé sous la forme proposée par Yates et Grundy (1953).

Lors de sa présentation de la méthode du partage des poids dans le contexte des enquêtes longitudinales, Ernst (1989) a proposé d'utiliser des constantes pour définir les poids d'estimation. Dans le contexte général de la MGPP, nous pouvons proposer l'utilisation du même genre de constantes. Définissons $\alpha_{j,ik} \geq 0$ pour toutes les paires

3. LA MÉTHODE GÉNÉRALISÉE DU PARTAGE DES POIDS

Décrite par Lavalée (1995), la MGPP est une généralisation de la méthode de partage des poids présentée par Ernst (1989) dans le contexte des enquêtes-ménages longs-tudinales. Gailly et Lavalée (1993) ont exposé diverses implications du recours à la méthode du partage des poids dans le cas de ces enquêtes. La MGPP peut être considérée comme une généralisation de l'échantillonnage en réservoir et aussi de l'échantillonnage en grappes adapté. Ces deux méthodes d'échantillonnage sont décrites par Thompson (1992) et par Thompson et Seber (1996).

Supposons qu'on sélectionne un échantillon s^A contenant m^A unités dans la population U^A contenant M^A unités selon un plan d'échantillonnage donné. Représentons par π_j^A la probabilité de sélection de l'unité j . Nous supposons que $\pi_j^A > 0$ pour toutes les unités $j \in U^A$.

Posons que la population U^B contient M^B unités. Cette population est répartie en N grappes, où la grappe i contient M_i^B unités. Par exemple, dans le contexte des enquêtes sociales, les grappes peuvent être des ménages et les unités, des personnes dans les ménages. Dans le cas des enquêtes-entreprises, les grappes peuvent être des entreprises et les unités, des établissements dans les entreprises. Dans notre exemple des exploitations agricoles, les grappes peuvent être des ménages, et les unités, des personnes dans les ménages qui produisent une déclaration de revenus pour l'ADRC.

Supposons qu'il existe un lien entre les unités j de la population U^A et les unités k de la grappe i de la population U^B . Ce lien est précisé par une variable indicatrice $l_{j,k}$, où $l_{j,k} = 1$ s'il existe un lien entre l'unité $j \in U^A$ et l'unité $k \in U^B$, et 0 autrement. Notons qu'il pourrait exister dans la population U^A certaines unités j pour lesquelles il n'existe aucun lien avec une unité k d'une grappe i de la population U^B , c'est-à-dire $L_j^A = \sum_{k=1}^{M_i^B} l_{j,k} = 0$ pour toutes les unités $j \in U^A$. En outre, il pourrait exister zéro, un ou plusieurs liens pour toute unité k d'une grappe i de la population U^B , c'est-à-dire $L_k^B = \sum_{j=1}^{M_i^A} l_{j,k} = 0$, $L_k^B = 1$ ou $L_k^B > 1$ pour toute unité $k \in U^B$.

Dans le cas de la MGPP, nous énonçons la contrainte qui suit :

Chaque grappe i de U^B doit posséder au moins un lien avec une unité j de U^A , c'est-à-dire $L_i^B = \sum_{j=1}^{M_i^A} l_{j,k} > 0$.

Il est essentiel d'imposer cette contrainte pour que la MGPP produise des estimations non biaisées. Nous verrons à la section 4 que, dans le contexte du couplage d'enregistrements, cette contrainte pourrait ne pas être satisfait. Pour chaque unité j sélectionnée dans s^A , nous repérons les unités ik de U^B qui donnent un appartenance de poids non nul avec j , c'est-à-dire $l_{j,ik} = 1$. Pour chaque unité ik repérée, nous supposons que nous pouvons établir la liste

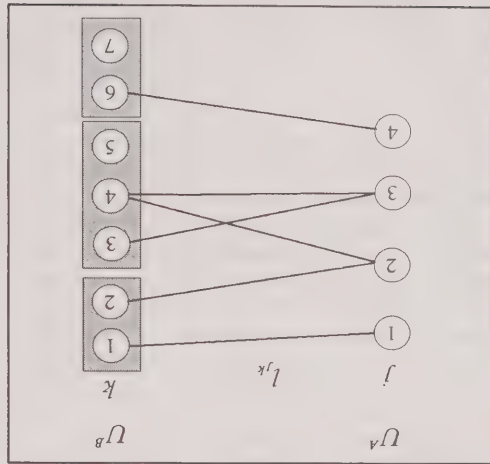


Figure 1. Exemple d'appariements

représentent la même population (au moyen d'un ensemble différent de caractéristiques), il est probable qu'à chaque unité j du fichier A ne soit couplée qu'une seule unité k du fichier B. Autrement dit, le mode de couplage sera binitivoque (plusieurs à un), co-unitivoque (un à plusieurs) ou multivoque (plusieurs à plusieurs). Comme nous l'avons mentionné plus haut, étant donné le caractère probabiliste du processus de couplage des enregistrements, lequel est susceptible d'introduire certaines erreurs, il pourrait exister plus d'un appartement par unité. En pratique, la résolution du problème nécessite habituellement une intervention manuelle. Dans l'exemple des exploitations agricoles, plusieurs exploitants d'une même exploitation pourraient présenter chacun une déclaration de revenus à l'ADRC (plusieurs). Pareillement, un agriculteur qui exploite plus d'une ferme pourrait ne présenter qu'une seule déclaration de revenus pour l'ensemble de ses opérations (correspondance multi-unitivoque, plusieurs à une). Enfin, on pourrait imaginer un scénario de liens multivoques (plusieurs à plusieurs) où un agriculteur exploite plus d'une entreprise agricole et où chacune de ces entreprises compte plusieurs exploitants distincts. Ces diverses situations sont représentées à la figure 1. Dans cette dernière, l'unité $j=1$ de U^A possède un lien un à un avec l'unité $k=1$ de U^B , l'unité $j=2$ forme un lien un à plusieurs avec les unités $k=2$ et $k=4$ et les unités $j=2$ et $j=3$ forment ensemble un lien plusieurs à un avec l'unité $k=4$. Il est évident qu'il est plus difficile de décider de la validité des appariements si l'on considère l'exemple des exploitations agricoles que celui d'une même population, puisque, dans le premier cas, il est possible d'observer de vrais appariements multivoques ou co-unitivoques.

fait souvent défaut et le couplage doit alors se faire selon une méthode probabiliste permettant de décider si deux enregistrements provenant, chacun, de l'un des fichiers forment ou non un appartement vrai. Selon cette méthode, on calcule la probabilité d'obtenir un appartement vrai, puis, d'après la valeur de celle-ci, on décide si les enregistrements forment vraiment un appartement.

De façon formelle, nous considérons la matrice $A \times B$ provenant des deux fichiers A et B. Représentons par j un enregistrement (ou unité) provenant du fichier A (ou de la population U^A) et par k un enregistrement (ou unité) provenant du fichier B (ou de la population U^B). Pour chaque paire (j, k) de $A \times B$, nous calculons un poids de couplage qui reflète la mesure dans laquelle il est probable que la paire (j, k) soit un appartement vrai. La probabilité que la paire (j, k) soit un appartement vrai est d'autant plus forte que le poids de couplage est élevé. Ordinairement, le poids de couplage se fonde sur le ratio des probabilités conditionnelles d'avoir un appartement μ et un non-appartement $\bar{\mu}$ étant donné le résultat de la comparaison C_{qjk} de la caractéristique q des enregistrements j provenant de A et des enregistrements k provenant de B, $q = 1, \dots, Q$. Autrement dit,

$$\theta_{jk} = \log_2 \left\{ \frac{P(\mu_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})}{P(\bar{\mu}_{jk} | C_{1jk} C_{2jk} \dots C_{Qjk})} \right\}$$

$$= \theta_{1jk} + \theta_{2jk} + \dots + \theta_{Qjk} + \theta_{*jk} \quad (2.1)$$

$$\text{où } \theta_{qjk} = \log_2 \left\{ \frac{P(C_{qjk} | \mu_{jk})}{P(C_{qjk} | \bar{\mu}_{jk})} \right\} \text{ pour } q = 1, \dots, Q, \text{ et}$$

$$\theta_{*jk} = \log_2 \left\{ \frac{P(\mu_{jk})}{P(\bar{\mu}_{jk})} \right\}.$$

Le modèle mathématique proposé par Fellegi et Sunter (1969) tient compte des probabilités d'une erreur de couplage des unités j provenant de A et des unités k provenant de B. Le poids de couplage est alors défini comme

$$\theta_{FS}^{jk} = \sum_{q=1}^Q \theta_{qjk}$$

où

$$\theta_{FS}^{jk} = \begin{cases} \log_2 \text{ si } y \text{ a concordance pour la caractéristique } q \text{ pour la paire } (jk) \\ \log_2 ((1 - \eta_{qjk}) / (1 - \bar{\eta}_{qjk})) \text{ autrement} \end{cases}$$

avec $\eta_{FS}^{jk} = P(\text{concordance pour la caractéristique } q | \mu_{jk})$ et $\bar{\eta}_{qjk} = P(\text{concordance pour la caractéristique } q | \bar{\mu}_{jk})$. Notons que la définition de θ_{FS}^{jk} sous-entend que les Q comparaisons sont indépendantes.

Notons que le poids de couplage θ_{jk} est égal à 0 si la probabilité conditionnelle d'avoir un appartement μ est égale à 0. Autrement dit, nous obtenons $\theta_{jk} = 0$ lorsque la probabilité d'avoir un appartement vrai pour (j, ik) est nulle. Après avoir calculé le poids de couplage θ_{jk} pour chaque paire (j, k) de $A \times B$ nous devons déterminer si la valeur de ce poids est suffisamment élevée pour que la paire (j, k) soit considérée comme un appartement vrai. Habituellement, on applique une règle pour prendre cette décision. Conformément à la méthode de Fellegi et Sunter, nous utilisons un seuil supérieur θ_{High} et un seuil inférieur θ_{Low} . La

$$D(j, k) = \begin{cases} \text{appartement si } \theta_{jk} \geq \theta_{High} \\ \text{appartement potentiel si } \theta_{Low} < \theta_{jk} < \theta_{High} \\ \text{non-appartement si } \theta_{jk} \leq \theta_{Low} \end{cases} \quad (2.3)$$

Les seuils inférieur et supérieur θ_{Low} et θ_{High} sont déterminés d'après des bornes d'erreur *a priori* établies en tenant compte des faux appartements et des faux non-appartements. Lors de l'application de la règle de décision (2.3), une certaine intervention manuelle est nécessaire pour résoudre les cas pour lesquels le poids de couplage est compris entre les seuils inférieur et supérieur. Ces décisions sont généralement prises par examen des données et utilisation de données auxiliaires. Par exemple, pour les exploitations agricoles, on pourrait se servir de variables comme la date de naissance, l'adresse de voirie et le code postal, pour lesquelles des données figurent dans les deux fichiers sources. Le fait que le processus de couplage d'enregistrements soit automatisé et probabiliste pourrait donner lieu à certaines erreurs. Cette question a été examinée par plusieurs auteurs, notamment Bartlett et coll. (1993), Belin (1993) et Winkler (1995).

L'application de la règle de décision (2.3) aboutit à la définition d'une variable indicatrice $f_{jk} = 1$ si la paire (j, k) est considérée comme un appartement vrai et 0, autrement. Pour les paires dont le poids de couplage est compris entre les seuils inférieur et supérieur, une certaine intervention manuelle peut être nécessaire pour décider de la validité de l'appartement. Dans le cas où les fichiers A et B

Le problème étudié ici consiste à estimer le total d'une caractéristique d'une population cible qui est naturellement divisée en grappes. En supposant que l'on obtienne l'échantillon par tirage d'unités dans les grappes, si l'on sélectionne au moins une unité d'une grappe, alors la grappe entière sera interviewée. Généralement, cette méthode permet de réduire les coûts et, éventuellement, de produire des estimations et des caractéristiques pour les grappes ainsi que pour les unités.

Nous essayons, dans le présent article, de répondre aux questions qui suivent.

- a) Pouvons-nous utiliser la MGPP pour résoudre le problème de l'estimation dans le cas de populations reliées par couplage d'enregistrements?
- b) Pouvons-nous adapter la MGPP de façon à tenir compte des poids de couplage dérivés du couplage d'enregistrements?
- c) La MGPP permettrait-elle de réduire le nombre de cas de résolution manuelle que demande le couplage d'enregistrements?
- d) S'il existe plus d'une façon d'appliquer la MGPP, l'une d'elles peut-elle être considérée comme étant la « meilleure »?

Nous montrons que la réponse est manifestement affirmative pour (a) et (b). Par contre, pour la question (c), nous constatons que le recours à la MGPP oblige à augmenter la taille de l'échantillon, donc qu'il augmente les coûts de collecte de données. En ce qui concerne la question (d), bien qu'il n'existe aucune réponse catégorique, certaines approches semblent généralement plus appropriées que d'autres.

Pour commencer, nous décrivons brièvement le couplage d'enregistrements. En deuxième lieu, nous décrivons la MGPP. En troisième lieu, nous adaptons la MGPP de façon à proposer trois méthodes distinctes qui tiennent compte des poids de couplage résultant du couplage d'enregistrements. Ces méthodes sont les suivantes : 1) utiliser tous les poids non nul en appliquant les poids de couplage respectifs, 2) utiliser tous les appartements au hasard par essai de Bernoulli. Pour chacune de ces méthodes, nous présentons l'estimateur non biaisé d'un total, ainsi qu'une formule de calcul de la variance. Enfin, pour comparer les trois méthodes proposées à la méthode classique (où la MGPP est appliquée à des appartements acceptés au moyen d'une règle de décision), nous présentons certains résultats de simulation.

2. LE COUPLAGE D'ENREGISTREMENTS

Les concepts du couplage d'enregistrements ont été introduits par Newcome, Kennedy, Axford and James (1959) et formalisés dans le modèle mathématique de

Fellegi et Sunter (1969). Comme l'ont décrit Bartlett, Krewski, Wang and Zielinski (1993), le couplage d'enregistrements est le processus qui consiste à regrouper au moins deux éléments d'information enregistrés séparément qui se rapportent à une même unité (personne ou entreprise). Le couplage d'enregistrements est parfois appelé appariement exact, par opposition à appariement statistique. Ce dernier procédé vise à coupler des fichiers qui ne comptent que quelques unités communes (consulter Budd et Radner 1969, Budd 1971, Okner 1972 et Singh, Malnet, Kinnack and Rowe 1993). Dans le cas de l'appariement statistique, le couplage se fonde sur des caractéristiques similaires plutôt que sur un identificateur unique. Nous nous limitons ici au contexte du couplage d'enregistrements. Mais nous pourrions aussi appliquer la théorie élaborée à l'appariement statistique.

Supposons que l'on dispose de deux fichiers A et B contenant des variables concernant deux populations U^A et U^B , respectivement et qu'il existe certains liens entre ces deux populations. Ainsi, il pourrait s'agir d'une même population pour laquelle chaque fichier contient des données sur un ensemble distinct de caractéristiques des unités de la population. Il pourrait aussi s'agir de deux populations différentes, mais présentant certains liens naturels. Par exemple, l'une d'elles pourrait être une population de parents et l'autre une population d'enfants appartenant à ces parents. Notons que les enfants vivent habituellement dans des ménages que l'on peut considérer comme des grappes. Un autre exemple est celui d'une enquête agricole, où la première population est une liste d'exploitations agricoles déterminée d'après le Recensement de l'agriculture du Canada et la deuxième, une liste de dossiers fiscaux produite par l'Agence des douanes et du revenu du Canada (ADRC). Dans la première population, chaque exploitation agricole est repérée par un identificateur unique appelé numéro d'identification de l'exploitation agricole et certaines variables supplémentaires, dont le nom et l'adresse des exploitants qui ont été fournis lors du recensement. La deuxième comprend les déclarations de revenus des particuliers qui ont déclaré un revenu agricole. Ces personnes vivent dans des ménages. L'identificateur unique qui figure dans ces dossiers est soit le numéro d'assurance sociale soit un numéro de société, selon que l'exploitation est, ou non, constituée en société. Cependant, chaque déclaration de revenus transmise à l'ADRC contient des variables similaires (nom et adresse du déclarant, etc.) à celles pour lesquelles des données sont recueillies dans le cadre du recensement.

L'objectif du couplage d'enregistrements est d'apparier les enregistrements de deux fichiers A et B. Si ces enregistrements contiennent un identificateur unique, alors le processus d'appariement est banal. Par exemple, dans l'exemple des exploitations agricoles, si les deux fichiers contiennent le numéro d'identification de l'exploitation, le couplage peut être exécuté selon une simple méthode d'appariement. Malheureusement, l'identificateur unique

Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements

PIERRE LAVALLÉE et PIERRE CARON¹

RÉSUMÉ

La combinaison de bases de données par des méthodes de couplage d'enregistrements en vue d'augmenter la quantité d'information disponible est un phénomène de plus en plus fréquent. Si l'on ne peut se fonder sur aucun identificateur unique pour procéder à l'appariement des enregistrements, on recourt au couplage probabiliste. On apparie un enregistrem-
ment du premier fichier à un enregistrement du deuxième avec une certaine probabilité et on décide ensuite si cette paire d'enregistrements représente ou non un appartement vrai. Habituellement, ce processus nécessite une certaine intervention manuelle qui demande du temps et des ressources humaines. En outre, il aboutit souvent à un couplage complexe. Autrement dit, au lieu d'être systématiquement biniivoque (un à un), le couplage entre les deux bases de données peut être multi-univoque (plusieurs à un), co-univoque (un à plusieurs) ou multivoque (plusieurs à plusieurs).

Le regroupement de deux bases de données par couplage d'enregistrements peut être considéré comme l'union de deux populations. Dans le présent article, nous étudions la production d'estimations concernant l'une des populations (la population cible) d'après un échantillon sélectionné dans l'autre population. Nous supposons que les deux populations ont été regroupées par couplage probabiliste d'enregistrements. Pour résoudre le problème d'estimation que pose le couplage complexe de la population dont est tiré l'échantillon à la population cible, Lavallée (1995) a proposé d'utiliser la méthode généralisée du partage des poids (MGPP) qui est une généralisation de la méthode du partage des poids présentée par Ernst (1989) dans le contexte des enquêtes longitudinales auprès des ménages.

Nous commençons par décrire brièvement le couplage d'enregistrements. En deuxième lieu, nous dérivons la MGPP. En troisième lieu, nous adaptons la MGPP afin de proposer trois méthodes distinctes qui tiennent compte des poids de couplage résultant du couplage des enregistrements. Ces méthodes sont les suivantes : 1) utiliser tous les appartements de poids non nul en appliquant les poids de couplage respectifs, 2) utiliser tous les appartements de poids non nul et supérieurs à un seuil établi et 3) choisir les appartements au hasard par essai de Bernoulli. Pour chacune de ces méthodes, nous présentons un estimateur non biaisé d'un total, ainsi qu'une formule de calcul de la variance. Enfin, nous présentons certains résultats de simulation afin de comparer les trois méthodes proposées à la méthode classique (où la MGPP est appliquée à des appartements acceptés au moyen d'une règle de décision).

MOTS CLÉS : Méthode généralisée du partage des poids; couplage d'enregistrements; estimation; grappes.

1. INTRODUCTION

Pour augmenter la quantité d'information disponible, il arrive de plus en plus fréquemment que l'on combine des données provenant de sources différentes. Souvent, les bases de données sont combinées par des méthodes de couplage d'enregistrements. Si les fichiers visés contiennent un identificateur unique utilisable, on réalise le couplage en se servant directement de cet identificateur comme clé d'appariement. En revanche, s'il n'existe aucun identificateur unique, on recourt à une méthode de couplage probabiliste. Dans ce cas, on détermine la probabilité qu'un enregistrement du premier fichier soit apparié à un enregistrement du second, puis on décide si la paire formée représente ou non un appartement vrai. Il convient de souligner qu'ordinairement, cette méthode nécessite, pour un certain nombre de cas, une résolution manuelle qui demande du temps et des ressources humaines.

Nous considérons ici la production de l'estimation d'un total (ou d'une moyenne) d'une population cible que l'on met en grappes en se servant d'un échantillon tiré d'une

autre population reliée à la première. Nous supposons que les deux populations sont reliées par couplage probabiliste d'enregistrements. Notons que cette méthode de couplage produit souvent des appartements complexes entre les deux populations. Autrement dit, le couplage entre les unités de chaque population n'est pas systématiquement biniivoque (un à un), et peut être multi-univoque (plusieurs à un), co-univoque (un à plusieurs) ou multivoque (plusieurs à plusieurs).

Pour résoudre le problème d'estimation que pose un couplage complexe entre la population dont est tiré l'échantillon et la population cible, Lavallée (1995) a proposé d'utiliser la méthode généralisée du partage des poids (MGPP) qui est une généralisation de la méthode de partage des poids présentée par Ernst (1989). Bien que cette dernière ait été mise au point dans le contexte d'enquêtes-ménages longitudinales, on a montré que la méthode du partage des poids peut être généralisée aux situations où une population cible de grappes est échantillonnée à l'aide d'une base de sondage représentant une population différente, mais liée d'une certaine façon à celle étudiée.

- BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A. et JOCELYN, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 4, 751-764.
- BREIDT, J., et FULLER, W.A. (1993). Regression weighting for multiphase samples. *Sankhya*, 55, 297-309.
- BREWSTER, K. (2000). Deriving and Estimating an Approximate Variance for the Horvitz-Thompson Estimator using only First Order Inclusion Probabilities. Dans *Proceedings of the Second International Conference on Establishment Surveys*. Buffalo, New York, 1417-1422.
- CASSADY, R.J., et VALLIANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.
- CHAUDHURI, A., et ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.
- COCHRAN, W.G. (1977). *Sampling Techniques*, 3-ième Ed. New York : John Wiley.
- DES RAJ (1968). *Sampling Theory*. TMH Edition.
- DEVILLE, J.-C. (1999). Estimation de la Variance pour des et de linéarisation. *Techniques d'enquêtes*, 25, 219-230.
- DEVILLE, J.-C. (1999). Calage simultané de plusieurs enquêtes. *Recueil du Symposium 1999 : Combiner des données de sources différentes*, 225-230.
- HARTLEY, H.O., et RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- HIDIROGLOU, M.A. (1995). Sampling and Estimation For Stage One of The Canadian Survey of Employment, Payrolls and Hours Survey. *Recueil de la Section des méthodes d'enquête*, Société Statistique du Canada, 123-128.
- HIDIROGLOU, M.A., LATOUCHE, M., ARMSTRONG, B. et GOSSIN, M. (1995). Improving Survey Information Using Administrative Records : The Case of the Canadian Employment Survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.
- KORN, E.L., et GRAUBARD, B.I. (1999). Analysis of Health Surveys. Wiley series in *Probability and Statistics*.
- MONTANARI, G.E. (1998). Estimation de la moyenne d'une large-scale surveys. *International Statistical Review*, 55, 191-202.
- MONTANARI, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.
- MONTANARI, G.E. (1997). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.
- MONTANARI, G.E. (1998). Estimation de la moyenne d'une population finie par régression. *Techniques d'enquête*, 24, 71-79.
- MONTANARI, G.E. (2000). Conditioning on auxiliary variables means in finite population inference. *Australian New Zealand Journal of Statistics*, 42, 407-421.
- NEWMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- RANCOURT, E., et HIDIROGLOU, M.A. (1998). L'utilisation de dossiers administratifs dans l'Enquête canadienne sur l'emploi, la rémunération, et les heures. *Recueil de la Section des Méthodes d'enquêtes*, Société Statistique du Canada, 39-48.
- RAO, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika*, 60, 125-133.
- RAO, J.N.K. (1994). Estimation of totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-166.
- RÖSEN, B. (2000). A User's guide to Pareto rps sampling. Dans *Proceedings of the Second International Conference on Establishment Surveys*. Buffalo, New York, 289-294.
- SÄRNDAAL, C.E. (1996). Efficient estimators with simple variances in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAAL, C.-E., SWENSSON, B. et WRETTMAN, Y. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- TAM, S. M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.
- TILLÉ, Y. (2001). *Théorie des Sondages : Échantillonnage et estimation en population finies*. Dumond.

Les moyennes qui leur correspondent sont

$$\bar{\varepsilon}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \varepsilon_{1k} \varepsilon_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \varepsilon_{2k}$$

et

$$\bar{\varepsilon}_{2h} = \frac{1}{n_{2h}} \sum_{k=1}^{n_{2h}} \varepsilon_{2k}$$

Ici, n_{2hg} est le nombre d'unités sélectionnées dans l'échantillon s_2 qui appartiennent à l'intersection des strates

U_h et s_{1g} .

Les résidus qui rentrent en jeu dans chacune de ces

expressions sont $\tilde{\varepsilon}_{1k} = g_{1k}(\gamma_k - x_{1k}' B_{1,OPT})$ et $\tilde{\varepsilon}_{2k} = g_{2k}(\gamma_k - x_{2k}' B_{2,OPT})$. Les facteurs d'ajustement g_{1k} et g_{2k} sont tels que définis dans la section 4.1.

L'Enquête sur l'emploi, la rémunération et les heures :

L'objectif de cette enquête est d'obtenir des estimations sur le nombre d'emplois rémunérés, la rémunération hebdomadaire moyenne et autres variables connexes, selon diverses combinaisons de branche d'activité et province. Cette enquête a été remaniée afin de permettre l'utilisation de dossiers administratifs sur toutes les entreprises comprises dans l'univers de l'enquête; ainsi, l'enquête produit des estimations basées à la fois sur les dossiers administratifs (échantillon ADMIN) et les données recueillies sur le terrain par l'Enquête sur la rémunération auprès des entreprises (ERE). Plus de détails sur l'ERE sont donnés dans Rancourt et Hidiroglou (1998).

L'échantillon ADMIN s_1 de quelques 200 000 unités est sélectionné de l'univers de retenue sur la source administrative U_1 des (comptes de retenue sur la paye) pour obtenir les données administratives. Le plan de sondage pour cet échantillon est Bernoulli stratifié (par région), et le taux d'échantillonnage dans les strates varie de 10 % à 100 %. La taille de l'échantillon représente environ 20% du nombre total de comptes de retenue sur la paye. Seulement deux variables $x_{1k}^{(1)}$ sont disponibles sur la source administrative : le nombre d'emplois rémunérés et la rémunération mensuelle brute.

L'échantillon de l'ERE, s_2 , est constitué d'environ 10 000 établissements sélectionnés à partir du Registre des entreprises U_2 . L'ERE recueille les mêmes deux variables que la source administrative soient le nombre d'employés rémunérés et la rémunération mensuelle brute, $x_{1k}^{(2)}$, plusieurs autres variables $x_{2k}^{(2)}$ d'intérêt définies par type d'employés (employés payés à l'heure, salariés, professionnels actifs, autres employés), ainsi que les variables d'intérêt telles le nombre d'heures payées et les gains hebdomadaires, $y_k^{(2)}$. On trouve plus d'information sur l'ERE dans Rancourt et Hidiroglou (1998).

L'ERE est stratifiée par type d'industrie, par région géographique et par taille (variant de deux à trois groupes définis selon le nombre d'employés). Ces strates ont été

6. CONCLUSION

Une seule formulation décrit les cas imbriqués et non-imbriqués. Ces deux cas sont habituellement traités séparément dans la littérature. Cet article a unifié ces deux méthodes de sondage en se basant sur la régression optimale. Pour le cas imbriqué, on a observé que l'estimateur de régression obtenu par Hidiroglou et Sándal (1998) ressemble beaucoup à la forme correspondante de l'estimateur de régression optimale. Pour le cas non-imbriqué, on a adapté la méthode de Deville (1999) lorsqu'il y a des données auxiliaires au niveau de la population. Finalement, on a illustré cette théorie par des exemples pratiques.

La seule formulation décrit les cas imbriqués et non-imbriqués. Ces deux cas sont habituellement traités séparément dans la littérature. Cet article a unifié ces deux méthodes de sondage en se basant sur la régression optimale. Pour le cas imbriqué, on a observé que l'estimateur de régression obtenu par Hidiroglou et Sándal (1998) ressemble beaucoup à la forme correspondante de l'estimateur de régression optimale. Pour le cas non-imbriqué, on a adapté la méthode de Deville (1999) lorsqu'il y a des données auxiliaires au niveau de la population. Finalement, on a illustré cette théorie par des exemples pratiques.

ou $w_k^{(2)}$ est le poids de sondage pour chaque établissement sélectionné, et $\hat{\sigma}_k^2$ sont des facteurs positifs connus qui contrôlent l'impact des données aberrantes ou le type d'estimateur. Donc, si $\hat{\sigma}_k^2$ est proportionnel à une des composantes de $x_k^{(2)}$, on obtient l'estimateur par quotient. L'estimateur d'une variable y est donc $\hat{Y} = \sum_{j=1}^J \sum_{i=1}^{s_{1j}} w_k^{(1)} x_{ik}^{(1)} \hat{B}_{ji}$, où s_{1j} correspond aux groupes de calage utilisé pour $s_{2,j}$. L'EBRH se sert donc d'un plan de sondage à échantillonnage double non-imbriqué. Plus de détails sur le remaniement de l'EBRH sont disponibles dans Hidiroglou (1995), et Hidiroglou, Latouche, Armstrong and Gossen (1995).

5.3 Deux enquêtes à Statistique Canada

Plusieurs enquêtes à Statistique Canada se servent de l'échantillonnage double. On illustre les idées de cet article par deux enquêtes menées auprès des entreprises. Ces

(EMD), et l'Enquête sur l'Emploi, la rémunération, et les enquêtes sont l'Enquête sur les marchandises du détail (EMD), l'Enquête sur les marchandises de détail (EMD), se sert de l'échantillonnage double imputé, tandis que l'Enquête sur l'emploi, la rémunération et les heures (ERH) se sert de l'échantillonnage double non-imputé

L'enquête trimestrielle sur les marchandises de détail : L'objectif de l'Enquête trimestrielle sur les marchandises de détail (ETMD) est d'obtenir de l'information détaillée sur les ventes des marchandises de détail tous les trois mois. L'ETMD est un sous-échantillon de l'Enquête mensuelle du commerce de détail (EMCD), une enquête mensuelle. L'EMCD mesure principalement les ventes par groupe de commerce (regroupement de codes de classification type industriel (CTI) à trois ou quatre chiffres de 1980), par province et pour certaines régions métropolitaines de recensement.

ment (RMR). La population cible se compose des compagnies statistiques ayant des emplacements statistiques identifiés sur le Registre des entreprises qui sont actifs dans le commerce de détail. Environ 16 000 compagnies sont interviewées chaque mois. La population est stratifiée par province, territoire, certaines RMR et par groupe de commerce. L'EMCD est stratifiée en *H* strates, basées sur la taille

(2-3-groupes, la géographie (10 provinces, 2 territoires) et le type d'industrie (16 groupes principaux). Cet échantillon est re-stratifié indépendamment pour l'ETDC. La stratification de l'ETDC diffère celle de l'EMDC au niveau géographique, taille et industrie. Un sous échantillon est sélectionné en se servant de la « nouvelle » stratification de l'échantillon de l'EMCD. L'estimation de l'ETDC est basée sur un estimateur à double quotient qui se sert des données auxiliaires (les ventes de l'EMCD. L'unité d'échantillonnage de deuxième phase (ETDC) demeure la

dominant, c'est-à-dire ceux pour lesquels elle génère le plus de ventes. L'estimateur par rapport à deux phases est utilisé par l'EMCD, Binder, Babyak, Brodeur, Hidiogrou, et Joeeyn (2000) ont dérivé l'estimateur de variance qui tenait compte du plan de sondage ainsi que de la méthode d'estimation. Ils ont réduit les estimateurs de variance du total à de simples sommes des termes (résidus) appropriés

Les résultats de Binder *et al.* (2000) peuvent être adaptés pour l'estimateur par ratio.

de sondage de l'FTMC peut être formellement décrit comme suit. La population est stratifiée en H strates $U^h, h = 1, \dots, H$, et des échantillons aléatoires simples s^h et sans remise, de taille n^h , sont sélectionnées dans chaque strate U^h . La variable x^k est observée pour chaque unité appartenant à s^h . L'échantillon de première phase qui en résulte, $s_1^h = U^h = 1, s_1^{h^*}$, est ensuite stratifié en strates $s_{1g}^h, g = 1, \dots, G$. La stratification de s_1^h est indépendante de la stratification de l'univers U . Un échantillon aléatoire simple s_2^g de taille n_2^g est ensuite sélectionné de chaque strate $s_{1g}^h, g = 1, \dots, G$. On observe (y^k, x^k) , où $x^k = (x_1^k, x_2^k, \dots, x_K^k)$, pour chaque unité appartenant à l'échantillon $s_2^g = U^{g-1} s_2^{g-1}$ et $y^k = x_1^k \beta + \varepsilon_2^k$ tiennent pour s_1^h et s_2^g respectivement. Pour chacun de ces modèles $(\varepsilon_2^k, \sigma_2^2(\varepsilon_2^k))$ et $\varepsilon_2^k \sim (0, \sigma_2^2(\varepsilon_2^k))$ où ε_2^k et ε_1^k sont des facteurs positifs connus. Si $z_1^k \neq 1$ ou $z_2^k \neq 1$ pour tous les unités $k \in U$, on peut standardiser les données en les divisant soit par $\sqrt{z_1^k}$ ou $\sqrt{z_2^k}$. L'estimateur de régression optimal pour le total X qui en résulte est :

où les composantes de \hat{Y}_{OPT} ont été définies dans la section 3.1. La forme simplifiée (sans doubles sommes) de la variance de \hat{Y}_{OPT} est :

$$\frac{N_2^y}{S_2^y} \left(f^y - 1 \right) = \sum_H^y N_2^y = V_{\text{OPT}}^y$$

$$\frac{\partial z}{\partial \theta} \left(f - 1 \right) \sum_{G=1}^G +$$

$$N_2^h(1-f^{1h}u_2^{1h}) - (1-f^{1h}u_2^{1h})u^{1h} \sum_G^{1=g} \sum_H^{1=h} S_2^{2h} u^{2h}$$

où les variances sont définies par

$$\left\{ \begin{array}{l} \\ 2 \end{array} \right\} \left(\frac{1}{I} u^{y_1} - \frac{1}{2} \sum_{k=1}^g \sum_{n_g} \frac{u^{z_g}}{u^{1_g}} \right) = S^{y_1}_{z_2}$$

$$\hat{S}_2^{2h_8} = \frac{1}{\sum_{n_{2h_8}^k=1}^k 1 - e^{-e_{1(h_8)}^{1k}}} =$$

et

$$\hat{S}_{2g} = \frac{1}{n_{2g}} \sum_{k=1}^K \frac{1 - n_{2g}^{2k}}{2k} \left(\hat{e}_{2k} - \hat{e}_{2k} \right).$$

La variance de $\hat{Y}_{\text{SEP, REG}}$ est estimée comme étant la somme des composantes de chaque phase, c'est-à-dire $V_1(\hat{Y}_{\text{EXP}})$ et $V_2(\hat{Y}_{\text{SEP, REG}})$. La variance $V_2(\hat{Y}_{\text{SEP, REG}})$ est obtenue en substituant la variable Y_k par $e_k = g_k(Y_k - x_k^k \hat{\theta}_h)$ dans l'expression $V_2(\hat{Y}_{\text{EXP}})$. Ceci implique que la variance estimée de $\hat{Y}_{\text{SEP, REG}}$ est :

$$\hat{V}(\hat{Y}_{\text{SEP, REG}}) = \frac{n_1}{N^2(1-f_1)} \sum_{h=1}^H p^{1/h} \left[(1-a_h) S_{2yh}^2 + \frac{n_1}{n_1} \left(\bar{Y}_{2h} - \bar{Y}_{2, \text{st}} \right)^2 \right]$$

$$+ \sum_{h=1}^H \frac{n_{2h}}{N^2(1-f_{2h})} p^{1/h} S_{2yh}^2$$

où

$$S_{2yh}^2 = \sum (e_k - \bar{e}_h)^2 / (n_{2h} - 1)$$

et

$$S_{2yh}^2 = \frac{1}{n_{2h}} \sum s_{2h}^k (Y_k^k - \bar{Y}_{2h}^k)^2$$

5.2 Échantillons non-imbriqués

Ces deux exemples sont extraits de Des Raj (1968, pages 142-149). Nous nous en servirons afin d'illustrer les résultats des sections 3 et 4. On considère deux plans de sondage différents.

Tableau 3
Deux plans de sondage avec échantillonnage imbriqué et non-imbriqué

Sondage	Plan de sondage	Estimateur	Variance	Non-imbriqué
$N \rightarrow n_1$ (EASSR)	$n_1 \rightarrow n_2$ (EPTAR)	$\hat{Y}_{\text{EPTAR}} = \frac{1}{N} \sum_{s_1} \frac{n_1}{Y_i} n_{2p_i}$	$N^2 \frac{n_1}{(1-f_1)} S_2^2 + \frac{n_2}{V(\hat{Y})}$	$N^2 \frac{n_1}{(1-f_1)} R^2 S_2^2 + \frac{n_2}{V(\hat{Y})} \left[1 + \frac{n_1}{1} (1-f_1) S_2^2 \right]$
$N \rightarrow n_1$ (EASSR)	$n_1 \rightarrow n_2$ (EASSR)	$\hat{Y}_{\text{RAT}} = \frac{\sum_{s_2} \frac{n_1}{Y_i}}{\sum_{s_1} \frac{n_2}{X_i}} = \bar{X} \bar{R}$	$N^2 \frac{n_1}{(1-f_1)} \left(2 R S_{xy} - R^2 S_2^2 \right) + N^2 \frac{n_2}{(1-f_2)} S_2^{y-Rx}$	$N^2 \frac{n_1}{(1-f_1)} R^2 S_2^2 + N^2 \frac{n_2}{(1-f_2)} S_2^{y-Rx}$
Premier plan de sondage		Deuxième plan de sondage		

Pour le deuxième plan de sondage on suppose que les deux échantillons s_1 et s_2 ont été sélectionnés avec un plan de sondage aléatoire simple et sans remise. Encore une fois, on considère les cas imbriqués et non-imbriqués. On suppose que l'on observe l'observation auxiliaire x_i pour toute unité sélectionnée dans le premier échantillon s_1 . L'estimateur est $\hat{Y}_{\text{RAT}} = (N/n_1) x_i' (\sum_{s_1} y_i / \sum_{s_2} x_i) = \bar{X} \bar{R}$. Le tableau 3 résume ces deux plans de sondage, et présente les deux estimateurs ainsi que leurs variances estimées pour les cas imbriqués et non-imbriqués.

Les termes non définis dans le tableau 3 sont $P_i = x_i / \sum_{s_1} x_i$; $p_i = x_i / \sum_{s_1} x_i$; $S_{y-Rx}^2 = (N-1)^{-1} \sum_{s_1} (y_i - R x_i)^2$; $f_1 = n_1/N$, et $R = Y/X$.

On remarque à partir du tableau 3 est que les variances ne sont pas très différentes entre les cas imbriqués et non-imbriqués. Pour \hat{Y}_{EPTAR} , la variance sera plus petite pour le cas imbriqué (CV) de la variable y est plus petit que celui de la variable x . Pour \hat{Y}_{RAT} , la variance sera plus petite pour le cas imbriqué si $p \text{ CV}(y) < \text{CV}(x)$ où p est la corrélation entre y et x .

La variance de \hat{Y}^{OPT} est :

$$V(\hat{Y}^{\text{OPT}}) = V\left(\hat{Y}_{\text{HT}}^{\text{HT}} - \frac{1}{1+\alpha}(\alpha\hat{X}_{\hat{X}}^1\hat{B}_{1,\text{OPT}} + \hat{X}_{\hat{X}}\hat{B}_{\text{OPT}})\right) + \frac{1}{1+\alpha}V(\hat{X}_{\hat{X}}^1\hat{B}_{1,\text{OPT}}) + \frac{1}{1+\alpha}V(\hat{X}_{\hat{X}}\hat{B}_{\text{OPT}})$$

Résultat 3 : La variance estimée de \hat{Y}^{OPT} (\hat{Y}^{OPT}), défini par l'équation (4.8) est approximativement égale à :

$$+ 2\alpha(\hat{B}'_{\text{OPT}}V(\hat{X}_{\hat{X}})\hat{B}_{1,\text{OPT}} + \text{Cov}(\hat{X}_{\hat{X}}^1, \hat{X}_{\hat{X}})\hat{B}_{\text{OPT}}) \quad (4.11)$$

$$+ \hat{B}'_{\text{OPT}}V(\hat{X}_{\hat{X}})\hat{B}_{\text{OPT}}$$

$$- \frac{1}{1+\alpha} \left[\alpha \hat{B}'_{1,\text{OPT}}V(\hat{X}_{\hat{X}}^1)\hat{B}_{1,\text{OPT}} + \hat{B}'_{1,\text{OPT}}V(\hat{X}_{\hat{X}})\hat{B}_{\text{OPT}} \right] + 2\alpha(\hat{B}'_{\text{OPT}}V(\hat{X}_{\hat{X}})\hat{B}_{1,\text{OPT}} + \text{Cov}(\hat{X}_{\hat{X}}^1, \hat{X}_{\hat{X}})\hat{B}_{\text{OPT}}) \quad (4.12)$$

Le calcul du premier terme de (4.12) est basé sur les résidus $y_k - (\alpha\hat{X}_{\hat{X}}^1\hat{B}_{1,\text{OPT}} + \hat{X}_{\hat{X}}\hat{B}_{\text{OPT}})/(1+\alpha)$. Les autres termes de (4.12) sont assez simples à calculer, se basant principalement sur les variances estimées de $\hat{X}_{\hat{X}}^1$ et de $\hat{X}_{\hat{X}}$, ainsi que sur leurs covariances. On peut se servir de l'approximation de la variance, telle qu'énoncé par Tillé (2001) pour ce cas, et aussi l'appliquer aux covariances.

5. QUELQUES EXEMPLES SPÉCIFIQUES

Trois exemples traditionnels pour l'échantillonnage double sont présentés pour les deux cas (imbriqué et non-imbriqué). En plus, nous décrivons brièvement deux grandes enquêtes entreprises menées à Statistique Canada qui se servent de ces méthodes.

5.1 Échantillons imbriqués

Exemple 1 : Supposons qu'un échantillon s_1 aléatoire simple de taille n_1 est sélectionné de la population U de taille N . L'échantillon est ensuite stratifié en L strates s_{1h} de taille respective n_{1h} . Des échantillons aléatoires s_{2h} de taille n_{2h} sont ensuite sélectionnés sans remise dans chaque strate s_{1h} . L'estimateur du total est $\hat{Y}^{\text{EXP}} = N \sum_{h=1}^L p_{1h} \bar{y}_{2h} = N \bar{y}_{2, \text{st}}$, où $p_{1h} = n_{1h}/n_1$. En se servant de (4.7), on peut démontrer que la variance estimée de \hat{Y}^{EXP} , $V(\hat{Y}^{\text{EXP}})$, se décompose en la somme de, $V_1(\hat{Y}^{\text{EXP}})$ et $V_2(\hat{Y}^{\text{EXP}})$ qui correspondent aux phases 1 et 2 du plan de sondage. Ainsi :

$$\hat{B}_{2h} = \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} x_k x_k' \right)^{-1} \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} x_k y_k \right)$$

si $k \in s_{2h}$. Pour chaque strate h , les pentes \hat{B}_h sont estimées par

$$g_{2k} = 1 + \left(\sum_{k \in s_{1h}} x_k x_k' - \sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} x_k x_k' \right)^{-1} \left(\sum_{k \in s_{2h}} \frac{n_{1h}}{n_{2h}} x_k y_k \right)$$

où

$$\hat{Y}^{\text{SEP, REG}} = \sum_{h=1}^L \frac{n_1}{N} \sum_{k \in s_{2h}} g_{2k} y_k$$

par régression séparée, c'est-à-dire,

Exemple 2 : Supposons que pour le plan de sondage de l'exemple 1, nous avons aussi des données auxiliaires, x_k , disponibles à la première phase s_1 . Si nous supposons que les pentes (\hat{B}_h) diffèrent entre les strates, nous posons le modèle $y_k = x_k' \hat{B}_h + \epsilon_k$, où $E(\epsilon_k) = 0$, $E(\epsilon_k^2) = \sigma_k^2$, $k \in s_{1h}$, $h = 1, \dots, L$, et $E(\epsilon_k \epsilon_{k'}) = 0$ pour $k \neq k'$, où $k, k' \in s_{1h}$, $h = 1, \dots, L$. Ce modèle nous amène à l'estimateur

$$\bar{y}_{2, \text{st}} = \sum_{h=1}^L p_{1h} \bar{y}_{2h}$$

et

$$\bar{y}_{2h} = \frac{1}{n_{2h}} \sum_{k \in s_{2h}} y_k$$

$$S_{2h}^2 = \frac{1}{n_{2h} - 1} \sum_{k \in s_{2h}} (y_k - \bar{y}_{2h})^2; \quad a_h = \frac{(n_1 - n_{1h})}{n_1}; f_{1h} = \frac{n_{1h}}{n_1}; f_{2h} = \frac{n_{2h}}{n_{1h}};$$

et

$$\hat{V}_2(\hat{Y}^{\text{EXP}}) = N^2 \sum_{h=1}^L \frac{n_{2h}}{(1 - f_{2h})} p_{1h} S_{2h}^2;$$

$$+ \left[\frac{n_1 - 1}{n_1} (\bar{y}_{2h} - \bar{y}_{2, \text{st}})^2 \right];$$

$$V_1(\hat{Y}^{\text{EXP}}) = N^2 \sum_{h=1}^L p_{1h} \left[(1 - a_h) S_{2h}^2 \right]$$

où

$$V(\hat{Y}^{\text{EXP}}) = V_1(\hat{Y}^{\text{EXP}}) + V_2(\hat{Y}^{\text{EXP}})$$

4. ESTIMATION DE LA VARIANCE DE L'ESTIMATEUR OPTIMALE DE REGRESSION

4.1 L'échantillonnage double imbriqué

Rappelons que l'estimateur de régression optimal de Y est

$$\hat{Y}^{\text{OPT}} = \hat{Y}^{\text{HT}} + (X_1 - \hat{X}_1)' \tilde{B}_{1,\text{OPT}} + (\hat{X} - \hat{X}_1)' \tilde{B}^{\text{OPT}} \quad (4.1)$$

Afin d'obtenir la variance estimée de (4.1), nous réécrivons les composantes définissant \tilde{B}^{OPT} ainsi que $\tilde{B}_{1,\text{OPT}}$ en regroupant les termes associés à la variable Y afin d'exprimer la double somme comme une simple somme. Cette manipulation a été décrite par Montanari (1998) pour un plan de sondage à une phase, ayant des probabilités de sélection arbitraires. En suivant Montanari (1998) et en adaptant l'algèbre du cas d'une phase à celui de deux phases on obtient :

$$\tilde{B}^{\text{OPT}} = \left[\sum_{s_1} \sum_{s_2} \sum_{\ell} \hat{c}_{1k\ell} x_k x_{\ell}' \right]^{-1} \left[\sum_{s_1} \sum_{\ell} \hat{c}_{1k\ell} x_k x_{\ell}' \right]$$

$$= \left[\sum_{s_1} \sum_{s_2} \sum_{\ell} \hat{c}_{1k\ell} x_k x_{\ell}' \right]^{-1} \left[\sum_{s_2} \sum_{\ell} \hat{a}_{2k}^* \frac{\pi_k^*}{Y_k} \right] \quad (4.2)$$

où

$$\hat{a}_{2k} = \frac{\pi_k^*}{1 - \pi_{2k|s_1}} x_k + \sum_{\ell \neq k}^{(1,2)} \frac{\pi_{2k\ell|s_1} \pi_{2\ell|s_1}^*}{\left(\pi_{2k\ell|s_1} - \pi_{2k|s_1} \pi_{2\ell|s_1} \right)} x_{\ell}.$$

On approxime $\tilde{B}_{1,\text{OPT}}$ donné en (3.15) par $[V(\hat{X}_1)]^{-1} [\text{Cov}(\hat{X}_1, \hat{Y}^{\text{HT}})]$, c'est-à-dire,

$$\tilde{B}_{1,\text{OPT}} = [V(\hat{X}_1)]^{-1} [\text{Cov}(\hat{X}_1, \hat{Y}^{\text{HT}})]$$

$$= \left[\sum_{s_1} \sum_{\ell} \hat{c}_{1k\ell} x_{1\ell} x_{1\ell}' \right]^{-1} \left[\sum_{s_1} \sum_{\ell} \hat{a}_{1k}^* \frac{\pi_{1\ell}}{Y_k} \right]$$

où

$$\hat{a}_{1k} = \frac{1 - \pi_{1k}}{\pi_{1k}} x_{1k} + \sum_{\ell \neq k}^{(1)} \frac{\pi_{1k\ell} \pi_{1\ell}^*}{\left(\pi_{1k\ell} - \pi_{1k} \pi_{1\ell} \right)} x_{1\ell}.$$

En substituant (4.2) et (4.3) dans (4.1), et en soustrayant le total de la population Y , on obtient :

$$\hat{Y}^{\text{OPT}} - Y = \left(\sum_{s_1} \sum_{Y_k} \hat{g}_{1k}^* \frac{\pi_{1k}}{Y_k} - \sum_{Y_k} \frac{\pi_k^*}{Y_k} \right) + \left(\sum_{s_2} \sum_{Y_k} \hat{g}_{2k}^* \frac{\pi_k^*}{Y_k} - \sum_{Y_k} \frac{\pi_k^*}{Y_k} \right) \quad (4.4)$$

où

$$\hat{g}_{1k} = 1 + (X_1 - \hat{X}_1)' V(\hat{X}_1)^{-1} a_{1k} \quad \text{pour } k \in s_1 \quad (4.5)$$

et

$$\hat{g}_{2k} = 1 + (X - \hat{X})' V(\hat{X})^{-1} a_{2k} \quad \text{pour } k \in s_2. \quad (4.6)$$

Résultat 2 : La variance estimée de \hat{Y}^{OPT} définie par l'équation (4.1) est :

$$V(\hat{Y}^{\text{OPT}}) = \sum_{s_2} \sum_{\ell} \hat{c}_{1k\ell} \hat{g}_{1k} \hat{g}_{1\ell} e_{1k} e_{1\ell} + \sum_{s_2} \sum_{\ell} \hat{c}_{2k\ell}^* \hat{g}_{2k} \hat{g}_{2\ell} e_{2k} e_{2\ell} \quad (4.7)$$

où

$$c_{1k}^* = \frac{\pi_{1k\ell} \pi_{1\ell}^*}{\left(\pi_{1k\ell} - \pi_{1k} \pi_{1\ell} \right)}; \quad c_{2k}^* = \frac{\pi_{2k\ell|s_1} \pi_{2\ell|s_1}^*}{\left(\pi_{2k\ell|s_1} - \pi_{2k|s_1} \pi_{2\ell|s_1} \right)};$$

$$e_{1k} = Y_k - x_{1k}' \tilde{B}_{1,\text{OPT}};$$

$$e_{2k} = Y_k - x_k' \tilde{B}^{\text{OPT}}.$$

4.2 L'échantillonnage double non-imbriqué

Nous obtenons la variance estimée de \hat{Y}^{OPT} en nous servant de l'approximation suivante.

$$\tilde{Y}^{\text{OPT}} = \hat{Y}^{\text{HT}} + (X_1 - \hat{X}_1)' \tilde{B}_{1,\text{OPT}} + (\hat{X} - \hat{X}_1)' \tilde{B}^{\text{OPT}} = \hat{Y}^{\text{OPT}} + O_p(n_1^{-1/2}) \quad (4.8)$$

où

$$\tilde{Y}^{\text{OPT}} = \hat{Y}^{\text{HT}} + (X_1 - \hat{X}_1)' \tilde{B}_{1,\text{OPT}} + (\hat{X} - \hat{X}_1)' \tilde{B}^{\text{OPT}}. \quad (4.9)$$

La décomposition de \tilde{Y}^{OPT} en termes plus élémentaires donne :

$$\tilde{Y}^{\text{OPT}} = \hat{Y}^{\text{HT}} + \left(X_1 - \frac{X_1 + \alpha \hat{X}_1}{\hat{X}_1 + \alpha \hat{X}_1} \right)' \tilde{B}_{1,\text{OPT}} + \left(\hat{X} - \frac{\hat{X} + \alpha \hat{X}}{\hat{X}} \right)' \tilde{B}^{\text{OPT}} + \left(\frac{1 + \alpha}{\hat{X}} - \frac{1 + \alpha}{1} \right) \alpha \hat{X}_1' \tilde{B}_{1,\text{OPT}} + \hat{X}_1' \tilde{B}_{1,\text{OPT}} + \hat{X}' \tilde{B}_{1,\text{OPT}} \quad (4.10)$$

optimale de α_2 est obtenue en minimisant la variance de \hat{X}_2 . Un choix sous optimal pour α_1 , est n_2 sont les tailles respectives des échantillons s_1 et s_2 . Il est à noter que Korn et Graubart (1999) ont aussi fait un tel choix pour combiner deux totaux étant estimés de deux sources différentes. En substituant \hat{X}_2 à la place de X_2 dans l'expression (3.25), on obtient

$$\hat{X}_2 - \hat{X}_2 = (\hat{X}_2 - \hat{X}_2) / (1 + \alpha_2). \quad (3.26)$$

L'estimateur de Y du total de la population est donc :

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\hat{X}_2 - \hat{X}_2)' \hat{B}_{2,\text{OPT}} \quad (3.27)$$

où

$$\hat{B}_{2,\text{OPT}} = - \left[V(\hat{X}_2 - \hat{X}_2) \right]^{-1} \text{Cov}(\hat{Y}_{\text{HT}}, (\hat{X}_2 - \hat{X}_2)') \quad (3.28)$$

Si l'on substitue (3.26) dans (3.28), on peut réexprimer $\hat{B}_{2,\text{OPT}}$ comme :

$$\hat{B}_{2,\text{OPT}} = \left[V(\hat{X}_2) \right]^{-1} \text{Cov}(\hat{Y}_{\text{HT}}, \hat{X}_2'). \quad (3.29)$$

Remarque : L'estimateur \hat{Y}_{OPT} (3.25) est exactement égal à \hat{Y}_{OPT} (3.27). Ceci implique qu'il n'y a eu aucun gain vis-à-vis de l'estimation de Y , en substituant un meilleur estimateur de X_2 . Cependant, l'estimateur $\hat{B}_{2,\text{OPT}}$ associé à \hat{Y}_{OPT} ressemble beaucoup plus à un estimateur de régression auquel on s'attend, que $\hat{B}_{2,\text{OPT}}$ associé à \hat{Y}_{OPT} . L'estimateur GREG correspond pour le cas où on sert de \hat{X}_2 au lieu de X_2 est :

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\hat{X}_2 - \hat{X}_2)' \hat{B}_{2,\text{GREG}} \quad (3.30)$$

où

$$\hat{B}_{2,\text{GREG}} = \left(\sum_{s_2} w_{2k} x_{(2)}^k / \sigma_{2k}^2 \right) \left(\sum_{s_2} w_{2k} x_{(2)}^k y_{(2)}^k / \sigma_{2k}^2 \right)^{-1}$$

En plus, si l'on connaît aussi $x_{(1)}^{1k}$ pour $k \in U_1$ où $X_1 = \sum_{s_1} x_{(1)}^{1k}$, on peut considérer l'estimateur de régression

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + (\hat{X}_1 - \hat{X}_1)' \hat{B}_{1,\text{OPT}} + (\hat{X}_2 - \hat{X}_2)' \hat{B}_{2,\text{OPT}} \quad (3.31)$$

C'est-à-dire que \hat{X} est obtenu en minimisant la combinaison linéaire $A\hat{X} + (I - A)\hat{X}$ et $V(\hat{X}) = \alpha V(\hat{X}_2)$. La différence entre \hat{X} et \hat{X} peut s'exprimer comme

$$\hat{X} - \hat{X} = (\hat{X} - \hat{X}) / (1 + \alpha). \quad (3.32)$$

Étant donné que s_1 et s_2 sont des échantillons indépendants, on peut aisément démontrer que :

$$\hat{B}_{\text{OPT}} = \left[V(\hat{X}) \right]^{-1} \text{Cov}(\hat{X}, \hat{Y}_{\text{HT}}') \quad (3.33)$$

et que

$$\hat{B}_{1,\text{OPT}} = \left[V(\hat{X}_1) \right]^{-1} \left[\text{Cov}(\hat{X}_1, \hat{Y}_{\text{HT}}') \right] \quad (3.34)$$

Les composantes de \hat{B}_{OPT} sont estimées par :

$$V(\hat{X}) = \sum_{s_2} \sum_{s_2} \hat{e}_{2k\ell} x_{(2)}^k x_{(2)}^\ell \quad (3.35)$$

et

$$\text{Cov}(\hat{X}, \hat{Y}_{\text{HT}}') = \sum_{s_2} \sum_{s_2} \hat{e}_{2k\ell} x_{(2)}^k y_{(2)}^\ell \quad (3.36)$$

tandis que celles de $\hat{B}_{1,\text{OPT}}$ sont estimées par :

$$V(\hat{X}_1) = \sum_{s_2} \sum_{s_2} \hat{e}_{2k\ell} x_{(2)}^{1k} x_{(2)}^{1\ell} \quad (3.37)$$

et

$$\text{Cov}(\hat{X}_1, \hat{Y}_{\text{HT}}') = \sum_{s_2} \sum_{s_2} \hat{e}_{2k\ell} x_{(2)}^{1k} y_{(2)}^\ell \quad (3.38)$$

$$\hat{e}_{2k\ell} = \frac{\pi_{2k\ell} (\pi_{2k} \pi_{2\ell})}{\pi_{2k\ell} - \pi_{2k} \pi_{2\ell}}.$$

où

$$\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (\hat{X}_1 - \hat{X}_1)' \hat{B}_{1,\text{GREG}} + (\hat{X}_2 - \hat{X}_2)' \hat{B}_{2,\text{GREG}} \quad (3.39)$$

où $X_1 = \sum_{s_1} x_{(1)}^{1k}$, $X_2 = \sum_{s_2} x_{(2)}^{2k}$ et $X = \sum_{s_1} x_{(1)}^{1k} + \sum_{s_2} x_{(2)}^{2k}$.

Les estimateurs de régression du type GREG dans l'équation (3.39) sont estimés par

$$\hat{B}_{1,\text{GREG}} = \left(\sum_{s_2} w_{2k} x_{(2)}^{1k} x_{(2)}^{1\ell} / \sigma_{2k}^2 \right) \left(\sum_{s_2} w_{2k} x_{(2)}^{1k} y_{(2)}^k / \sigma_{2k}^2 \right)^{-1} \quad (3.40)$$

et

$$\hat{B}_{2,\text{GREG}} = \left(\sum_{s_2} w_{2k} x_{(2)}^k x_{(2)}^k / \sigma_{2k}^2 \right) \left(\sum_{s_2} w_{2k} x_{(2)}^k y_{(2)}^k / \sigma_{2k}^2 \right)^{-1} \quad (3.41)$$

$$\hat{Y}_{\text{GREG}}^{\text{HT}} = \hat{Y}_{\text{HT}} + \left(X_1 - \hat{X}_1 \right)' \hat{B}_{1, \text{GREG}} + \left(X_{\hat{\hat{2}}} - \hat{X}_{\hat{\hat{2}}} \right)' \hat{B}_{\text{GREG}}$$

(3.21)

(3.21)

$$\mathbf{B}_{1, \text{GREG}} = \left(\sum_{s_1} \frac{w_{1k}^* x_{1k} x'_{1k}}{\sigma_{1k}^2} \right)^{-1} \left\{ \sum_{s_2} \frac{w_{1k}^* x_{1k} y_k}{\sigma_{1k}^2} + \sum_{s_2} \frac{w_{1k}^* x_{1k} x'_{1k}}{\sigma_{1k}^2} \mathbf{B}_{\text{GREG}} \right\} \quad (3.22)$$

où $\{\sigma_{1k}^z: k \in s_1\}$ et $\{\sigma_{2k}^z: k \in s_2\}$ sont des facteurs

où $\{\sigma_{1k}^z: k \in s_1\}$ et $\{\sigma_{2k}^z: k \in s_2\}$ sont des facteurs

(57.5)

(47°C)

$$\frac{2}{2} \quad \frac{2}{2} \quad \frac{2}{2}$$

ЭНЦИКЛОПЕДИЯ

5.2 Application à l'échantillonnage double non-impair

3.2 Application of the 1960s

La variance estimée de Y^{GREG} est

où $c_{2k\ell} | s_1 = c_{2k\ell} | s_1' / \pi_{2k\ell} | s_1'$.
 Rappelons que l'estimateur optimal par régression $B_{1,OPT}$ est donné par l'expression (3.9), où

$$T_1 = V(X_1)$$

$$H_1 = C\hat{O}V(X_1, Y_{HT}^{HT}) + C\hat{O}V(X_1, X_1') \Big| \hat{B}_{OPT}^{OPT} - C\hat{O}V(X_1, X_1') \Big| \hat{B}_{OPT}^{OPT}.$$

Nous définissons les composantes des termes T_1 et de H_1 comme suit. En premier lieu, nous estimons $V(X_1) = \sum_{s_1} c_{1k\ell} x_{1k} x_{1\ell}'$ par

$$V(X_1) = \sum_{s_1} \sum_{i\ell} c_{1k\ell} x_{1k} x_{1\ell}' \quad (3.14)$$

où $c_{1k\ell} = (\pi_{1k} \pi_{1\ell} - \pi_{1k} \pi_{1\ell}) / (\pi_{1k} \pi_{1\ell})$ et $c_{1k\ell} = c_{1k\ell} / \pi_{1k\ell}$. Aussi, puisque

$$C\hat{O}V(X_1, Y_{HT}^{HT}) = E_1 C\hat{O}V_2 \left[\left(X_1, Y_{HT}^{HT} \right) | s_1 \right] + C\hat{O}V_1 \left[E_2 \left(X_1 | s_1 \right), E_2 \left(Y_{HT}^{HT} | s_1 \right) \right]$$

$$= C\hat{O}V_1 \left(X_1, Y_{HT}^{HT} \right) = \sum_{U_1} c_{1k\ell} x_{1k} x_{1\ell}' \quad (3.15)$$

on estime $C\hat{O}V(X_1, Y_{HT}^{HT})$ par

$$C\hat{O}V(X_1, Y_{HT}^{HT}) = \sum_{s_2} \sum_{i\ell} c_{1k\ell} x_{1k} x_{1\ell}' \quad (3.16)$$

où

$$c_{1k\ell}^* = c_{1k\ell} / \pi_{k\ell}^*, \pi_{k\ell}^* = \pi_{1k\ell} \pi_{2k\ell} | s_1', \\ \pi_{1k\ell} = \Pr(k, \ell \in s_1), \\ \pi_{2k\ell} | s_1' = \Pr(k, \ell \in s_2 | s_1') \\ \text{et } \pi_{k\ell}^* = \pi_{1k} \pi_{2k} | s_1'.$$

De même,

$$C\hat{O}V(X_1, X_1') = \sum_{s_2} \sum_{i\ell} c_{1k\ell}^* x_{1k} x_{1\ell}' \quad (3.17)$$

$$C\hat{O}V(X_1, X_1') = \sum_{s_1} \sum_{i\ell} c_{1k\ell} x_{1k} x_{1\ell}' \quad (3.18)$$

Dans le cas de l'échantillonnage double imbriqué, l'estimateur optimal de B_1 est donc :

$$\hat{B}_{1,OPT} = \left(V(X_1) - [C\hat{O}V(X_1, Y_{HT}^{HT}) + C\hat{O}V(X_1, X_1') - C\hat{O}V(X_1, X_1') \Big| \hat{B}_{OPT}^{OPT}] \right) \hat{B}_{OPT}^{OPT} \quad (3.19)$$

le calcul de probabilités conjointes. Nous pouvons, cependant, tirer profit de la forme optimale, et l'exprimer de façon plus simple pour de nombreux plans de sondage. Pour des plans de sondage où la sélection de l'échantillon est avec probabilité inégale et sans remise, on peut contourner le calcul de probabilité conjointe par approximation de la variance exacte. Plusieurs auteurs, dont Hartley et Rao (1962), Deville (1999), Berger (1998), Rösen (2000), ainsi que Brewer (2000). Pour un échantillon s (tiré sans remise et à une phase), Tillé (2001) propose l'approximation suivante de la variance estimée de $Y_{HT}^{HT} = \sum_{s_1} y_k / \pi_k$, où π_k est la probabilité de sélection de l'unité k :

$$V(Y_{HT}^{HT}) = \sum_{s_1} \left(\frac{y_k^2}{c_k} (y_k - y_k^*)^2 \right) = \sum_{s_1} c_k \left(\frac{y_k^2}{y_k} - y_k^* \right)^2 \quad (3.20)$$

Ici, c_k est la variable qui sert à l'approximation, $y_k^* = \pi_k \sum_{s_1} c_{\ell} y_{\ell} / \pi_{\ell}^*, y_k^* = y_k / \pi_k$ et π_k^* est la probabilité de sélection de l'unité k . Tillé (2001) fournit plusieurs exemples de c_k pour divers plans de sondage. Notons que cette formule est exacte dans le cas d'un plan de sondage stratifié, où l'échantillon est aléatoire simple et sans remise dans chaque strate U_h ($h = 1, \dots, L$) de la population U . Pour ce cas, si k dénote une unité appartenant à l'échantillon s_h sélectionné dans la strate U_h , alors $c_k = n_h / N_h$ si $k \in U_h$ et 0 autrement, $\pi_k = n_h / N_h$ si $k \in U_h$ et 0 autrement. Nous obtenons ainsi la variance estimée exacte, $V = (Y_{HT}^{HT})^2 = \sum_{h=1}^L N_h^2 (1 - n_h / N_h) \sum_{s_h} (y_k - \bar{y}_h)^2 / n_h (n_h - 1)$. La formule est aussi exacte dans le cas d'un plan de sondage stratifié où l'échantillon a été sélectionné avec remise. Ici $c_k = 1$ pour toutes les unités appartenant à la strate U_h et zéro autrement. En ce servant de cette approximation, les doubles sommes qui apparaissent dans $B_{1,OPT}$ ainsi que $B_{1,OPT}$ peuvent être exprimées comme de simples sommes. Hidiroglou et Särndal (1998) ont contourné le problème de doubles sommes, dans l'estimation de B_1 et B_1 en posant l'estimateur GREG, \hat{Y}_{GREG} pour un plan de sondage à deux phases imbriquées :

En résolvant le système d'équations (3.3) et (3.4), nous obtenons les paramètres requis B et B_1 . C'est-à-dire :

$$(3.5) \quad B = T^{-1}H \quad \text{ou} \quad T = V(X - \hat{X}) - (\text{Cov}(X_1, (X - \hat{X})'))',$$

$$V^{-1}(X_1)(\text{Cov}(X_1, (X - \hat{X})'))',$$

$$H = (\text{Cov}(X_1, (X - \hat{X})), Y_{HT}') +$$

$$+ (\text{Cov}(X_1, (X - \hat{X})')' V^{-1}(X_1) \text{Cov}(X_1, Y_{HT}')$$

$$(3.6) \quad B_1 = T_1^{-1}H_1 \quad \text{et}$$

$$T_1 = V(X_1),$$

$$H_1 = \text{Cov}(X_1, Y_{HT}') + \text{Cov}(X_1, (X - \hat{X})')' B.$$

Résultat 1 : Un estimateur de régression optimal pour les échantillons imbriqués et non-imbriqués est :

$$(3.7) \quad Y_{\text{OPT}} = Y_{HT} + (X_1 - \hat{X})' B_{1,\text{OPT}} + (\hat{X} - \hat{X})' B_{\text{OPT}}$$

$$(3.8) \quad \hat{B}_{\text{OPT}} = \hat{T}^{-1} \hat{H}$$

$$(3.9) \quad \hat{B}_{1,\text{OPT}} = \hat{T}_1^{-1} \hat{H}_1.$$

$\hat{T}_1, \hat{H}_1, \hat{T}$ et \hat{H} sont les valeurs estimées de T_1, H_1, T et H obtenues à l'aide d'un cadre de travail menant à l'inférence basée sur le plan de sondage. Ces valeurs dépendent de la méthode de sélection des échantillons. La variance de Y_{OPT} et de son estimateur associé dépendent de l'imbication ou non du plan de sondage. Puisque l'estimateur Y_{OPT} est optimal, il s'ensuit que l'estimateur par régression Y_{OPT} est aussi optimal. Cette forme a été discutée par Montanari (1987) pour le cas d'un plan de sondage à une phase.

3.1 Application à l'échantillonnage double imbriqué

La théorie pour ce cas est développée en se servant d'une approche conditionnelle. Si deux paramètres quelconques θ_1 et θ_2 sont estimés par $\hat{\theta}_1$ et $\hat{\theta}_2$ à partir de l'échantillon s_2

(ii) La variance de $\hat{\theta}$ est :

$$(3.10) \quad V(\hat{\theta}) = E_1 V_2(\hat{\theta} | s_1) + V_1 E_2(\hat{\theta} | s_1).$$

(iii) La covariance entre $\hat{\theta}_1$ et $\hat{\theta}_2$ est :

$$\text{Cov}(\hat{\theta}_1, \hat{\theta}_2) = E_1 \text{Cov}_2((\hat{\theta}_1, \hat{\theta}_2) | s_1)$$

$$+ \text{Cov}_1(E_2(\hat{\theta}_1 | s_1), E_2(\hat{\theta}_2 | s_1)).$$

Les différentes composantes de \hat{T}, H, \hat{T}_1 et de \hat{H}_1 seront estimées en supposant un plan de sondage ayant une taille d'échantillon non-fixe. Le cas de la taille l'échantillon fixe suit facilement, étant un cas spécial de ce dernier. En se servant des expressions (i) – (iii), on peut réexprimer les termes qui définissent le paramètre B comme :

$$\text{Cov}(X, \hat{X}) = \text{Cov}(\hat{X}, X) = V(\hat{X});$$

$$\text{Cov}(Y_{HT}, X') = \text{Cov}(Y_{HT}, X');$$

$$V(X - \hat{X}) = E_1 \left[\sum_{s_1} \sum_{s_2 | s_1} C_{2k | s_1} X_k X_k' \right];$$

$$\text{Cov}[X_1, (X - \hat{X})'] = 0;$$

et

$$\text{Cov}(X, Y_{HT}) = \text{Cov}(X, Y_{HT})$$

$$(3.11) \quad + E_1 \left[\sum_{s_1} \sum_{s_2 | s_1} C_{2k | s_1} X_k X_k' \right];$$

où $C_{2k | s_1} = (\pi_{2k | s_1} - \pi_{2k | s_1} \pi_{1k}^*) / \pi_{1k}^* \pi_{1k}^*$ et $Y_{HT} = \sum_{s_1} Y_k \pi_{1k}^*$. Les probabilités d'inclusion dans ces expressions sont $\pi_{2k | s_1} = \text{Pr}(k, \ell \in s_2 | s_1)$ et $\pi_{1k}^* = \pi_{1k} \pi_{2k | s_1}$. On peut exprimer B plus simplement comme :

$$B = \left[E_1 \left(\sum_{s_1} \sum_{s_2 | s_1} C_{2k | s_1} X_k X_k' \right) \right]^{-1}$$

$$E_1 \left[\sum_{s_1} \sum_{s_2 | s_1} C_{2k | s_1} X_k Y_k \right]$$

et son estimateur optimal est :

$$(3.13) \quad \hat{B}_{\text{OPT}} = \left[\sum_{s_2} \sum_{s_1} C_{2k | s_1} X_k X_k' \right]^{-1} \left[\sum_{s_2} \sum_{s_1} C_{2k | s_1} X_k Y_k \right]$$

dans s_1 et s_2 serait fort probablement causé par une différence dans le questionnaire ou par le fait que différents répondants remplissent le questionnaire. Notons aussi que $X_1 = \sum_{i \in U_1} x_{1i}^{(1)} = \sum_{i \in U_2} x_{1i}^{(2)}$ puisque U_1 et U_2 ont la même couverture.

3. ESTIMATEUR OPTIMAL POUR LES ÉCHANTILLONS IMBRIQUÉS ET NON-IMBRIQUÉS

Pour les deux cas, imbriqué et non-imbriqué, l'objectif est d'estimer le total de population $Y = \sum_{i \in U} y_i$ où y_k représente la valeur de l'unité $k \in U$. Un estimateur non biaisé de Y est $\hat{Y}_{HT} = \sum_{i \in s_2} w_k^{(2)} y_k$, où $w_k^{(2)} = w_{1k} w_{2k}$ pour le cas imbriqué et $w_k^{(2)} = w_{2k}$ pour le cas non-imbriqué.

Les données auxiliaires nous permettent de modifier les poids d'échantillonnage et à l'aide de facteurs de calage calculés d'après l'information complémentaire à différents niveaux (univers, échantillon de première phase). On modifie le poids d'échantillonnage d'une unité en le multipliant par un facteur de calage, et on appelle le produit « poids de calage ». Le tableau 1 résume la représentation des données disponibles pour les cas imbriqués et non-imbriqués, correspondant aux figures 1 et 2.

Tableau 1

Données disponibles pour la population et les échantillons		
Niveau	Cas imbriqué	Cas non-imbriqué
Population	x_{1k} : connu pour $k \in U$	x_{1k} : connu pour $k \in U_1$
Premier échantillon	x_k : observé pour $k \in s_1$	x_k : observé pour $k \in s_1$
Deuxième échantillon	y_k, x_k : observé pour $k \in s_2$	$y_k, x_k^{(2)}$: observé pour $k \in s_2$

Un estimateur de régression qui peut servir à estimer le total Y de la population pour les échantillons imbriqués et non-imbriqués est :

et

$$V(\hat{X} - \hat{X})B + \text{Cov}(\hat{X} - \hat{X}, \hat{Y}_{HT}) \\ - \text{Cov}(\hat{X} - \hat{X}, \hat{X}_1')B_1 = 0 \quad (3.3)$$

Les paramètres sont obtenus en prenant la dérivée de (3.2) par rapport à B et B_1 . Nous obtenons les deux équations suivantes à résoudre :

$$V(\hat{Y}_{REG}) = V(\hat{Y}_{HT}) + B_1' V(\hat{X}_1)B_1 + B' V(\hat{X} - \hat{X})B \\ - 2 \text{Cov}(\hat{Y}_{HT}, \hat{X}_1')B_1 + 2 \text{Cov}(\hat{Y}_{HT}, (\hat{X} - \hat{X})')B \\ - 2B_1' \text{Cov}(\hat{X}_1, (\hat{X} - \hat{X})')B. \quad (3.2)$$

B_1 . Cette variance est : la variance de \hat{Y}_{REG} afin d'obtenir les estimateurs de B et pour les deux cas imbriqué et non-imbriqué. On minimise les covariances $\text{Cov}(\hat{X}_1, \hat{X}_1')$, $\text{Cov}(\hat{X}_1, \hat{X})$, $\text{Cov}(\hat{X}_1, \hat{X}')$, $\text{Cov}(\hat{Y}_{HT}, \hat{X}_1')$ et $\text{Cov}(\hat{Y}_{HT}, \hat{X})$ sont connues ou estimables.

2. On suppose aussi que les variances, $V(\hat{Y}_{HT})$, ainsi que les covariances $\text{Cov}(\hat{X}_1, \hat{X}_1')$, $\text{Cov}(\hat{X}_1, \hat{X})$, $\text{Cov}(\hat{X}_1, \hat{X}')$, $\text{Cov}(\hat{Y}_{HT}, \hat{X}_1')$ et $\text{Cov}(\hat{Y}_{HT}, \hat{X})$ sont connues ou estimables.

Les différentes valeurs des totaux des données auxiliaires x et y dans l'équation (3.1) sont présentées dans le tableau

$$\hat{Y}_{REG} = \hat{Y}_{HT} + (\hat{X}_1 - \hat{X}_1')B_1 + (\hat{X} - \hat{X})'B. \quad (3.1)$$

Tableau 2
Sommes des données auxiliaires x et y pour les cas imbriqués et non-imbriqués

Niveau	Cas imbriqué	Cas non-imbriqué
Population	$X_1 = \sum_{i \in U} x_{1i}$	$X_1 = \sum_{i \in U_1} x_{1i}$
Premier échantillon	$\hat{X}_1 = \sum_{i \in s_1} w_{1i} x_{1i}; \hat{X} = \sum_{i \in s_1} w_{1i} x_k$	$\hat{X}_1 = \sum_{i \in s_1} w_{1i} x_{1i}; \hat{X} = \sum_{i \in s_1} w_{1i} x_k$
Deuxième échantillon	$\hat{X}_1 = \sum_{i \in s_2} w_k^{(2)} x_{1i}; \hat{X} = \sum_{i \in s_2} w_k^{(2)} x_k$	$\hat{X}_1 = \sum_{i \in s_2} w_{2k} x_{1i}; \hat{X} = \sum_{i \in s_2} w_{2k} x_k$
		$\hat{Y}_{HT} = \sum_{i \in s_2} w_{2k}^{(2)} y_k$

2. NOTATION

2.1 Cas imbriqué

(1984) afin d'obtenir les probabilités conjointes requises pour le calcul de l'estimation de la variance pour un

estimateur de total $Y = \sum U_k Y_k$.

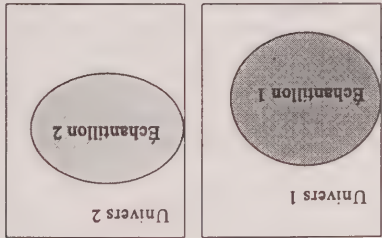


Figure 2. Deux échantillons indépendants bases de sondages différentes

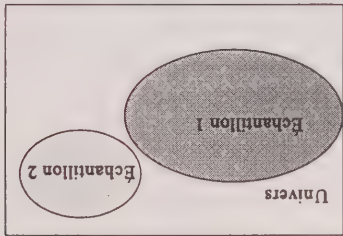


Figure 3. Deux échantillons sélectionnés indépendamment dans la même base de sondage

Pour le cas que nous allons étudier, nous supposons que les échantillons s_1 et s_2 ont été sélectionnés indépendamment à partir de deux bases de sondage différentes (Figure 2), $U_1 = \{1, \dots, k, \dots, N_1\}$ et $U_2 = \{1, \dots, k, \dots, N_2\}$. Les échantillons $s_1 (s_1 \subseteq U_1)$ et $s_2 (s_2 \subseteq U_2)$ sont sélectionnés tel que leur probabilité de sélection est respectivement $\pi_{1k}^{(1)} = P(k \in s_1) > 0$ et $\pi_{2k}^{(2)} = P(k \in s_2) > 0$. Le poids de l'unité k est $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ pour le premier échantillon s_1 et $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ pour le deuxième échantillon s_2 . Les indices supérieurs (1) et (2) sont utilisés pour différencier les probabilités de sélection du cas imbriqué. Il se peut que les unités de sondage diffèrent entre les deux bases, tout en ayant la même couverture. Des exemples de telles procédures d'échantillonnage ont été mentionnés dans l'introduction, et plus de détails sont donnés dans le deuxième exemple de la section 5.3.

Soit $x_k = (x_{1k}^{(1)}, x_{2k}^{(2)})$, un vecteur de données auxiliaires. On suppose que $x_{1k}^{(1)}$ est connue pour toutes les unités appartenant à la base de sondage U_1 , tandis que $x_{2k}^{(2)}$ est seulement observé pour l'échantillon s_1 . Nous observons $y_k^{(2)}, x_k^{(2)}$ pour l'échantillon s_2 . Le degré de divergence entre les valeurs des données varie selon la complexité de l'unité d'échantillonnage, et combien ces unités diffèrent en concept entre les deux bases. Pour les unités plus « simples », les données rapportées pour des unités « semblables » dans s_1 et s_2 devraient être à peu près égales. Une différence de similitude entre les mêmes unités

On suppose que $\pi_{1k}^{(1)} > 0$ pour toutes les valeurs $k \in U$ et que $\pi_{2k}^{(2)} > 0$ pour toutes les valeurs $k \in s_1$. Le poids de l'unité échantillonnée k sera dénoté par $w_{1k}^{(1)} = 1/\pi_{1k}^{(1)}$ pour l'échantillon de la première phase et $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ pour celui de la deuxième phase. Le poids d'échantillonnage global d'une unité de deuxième phase sélectionnée, $k \in s_2$, sera donc $w_k^* = w_{1k}^{(1)} w_{2k}^{(2)}$.

La population est représentée par $U = \{1, \dots, k, \dots, N\}$. On préleve un premier échantillon probabiliste $s_1 (s_1 \subseteq U)$ sélectionné de la population U selon un plan d'échantillonnage pour lequel la probabilité de sélection est $\pi_{1k}^{(1)} = P(k \in s_1)$ pour la $k^{\text{ième}}$ unité choisie dans s_1 . Étant donné s_1 un deuxième échantillon $s_2 (s_2 \subseteq s_1 \subseteq U)$ est tiré de s_1 selon un plan d'échantillonnage pour lequel la probabilité conditionnelle de sélection est $\pi_{2k|s_1}^{(2)} = P(k \in s_2 | s_1)$ pour la $k^{\text{ième}}$ unité choisie dans s_2 . Notons qu'il s'agit de probabilités conditionnelles puisqu'elles supposent qu'on connaît s_1 . La figure 1 représente un exemple d'échantillons imbriqués.

Appelons x le vecteur auxiliaire disponible de l'échantillon de la première phase, et x_k la valeur pour l'unité k . Comme le font Hidiroglou et Sæmstad (1998), divisons x_k en deux parties $x_{1k}^{(1)}$ et $x_{2k}^{(2)}$. Les valeurs du vecteur $x_{1k}^{(1)}$ sont supposées connues pour l'ensemble de la population U , alors que les valeurs du vecteur $x_{2k}^{(2)}$ ne sont disponibles que pour l'échantillon de la première phase s_1 .

2.2 Cas non-imbriqué

Il est possible que les deux échantillons soient sélectionnés indépendamment de la même base de sondage, ou même de bases de sondage différentes (mais équivalentes). Les figures 2 et 3 représentent des exemples de cas non-imbriqués.

Le cas non-imbriqué représenté par la figure 3 n'est pas considéré dans cet article. Ce cas peut être compliqué pour des plans de sondages arbitraires, puisqu'il faut calculer des probabilités de sélection conjointes entre les deux échantillons s_1 et s_2 . Ce calcul se simplifie si l'on considère que deux échantillons s_1 et s_2 ont été sélectionnés en se servant d'un plan de sondage tel que le plan aléatoire simple (avec

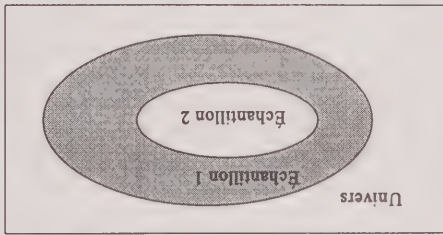


Figure 1. Échantillons imbriqués

L'échantillonnage double

M.A. HIDIROGLOU¹

RÉSUMÉ

La théorie de l'échantillonnage double est d'habitude présentée en supposant que l'un des échantillons est imbriqué dans l'autre. Ce genre de sondage est appelé sondage à deux phases. L'échantillon de première phase fournit de l'information auxiliaire (x) relativement peu chère à obtenir, alors que l'échantillon de deuxième phase contient les variables d'intérêt. On se sert des données de la première phase de plusieurs façons : (a) pour stratifier l'estimation en se servant de l'estimateur par différence, quotient ou régression; (c) pour sous échantillonner un ensemble d'unités non répondantes. Cependant, il n'est pas nécessaire que l'un des échantillons soit imbriqué dans l'autre ou soit sélectionné dans la même base de sondage. Le cas de l'échantillonnage double *non-imbriqué* est légèrement abordé dans les livres classiques du sondage (Des Raj 1968, Cochran 1977). Cette méthode est présentement utilisée dans plusieurs organismes nationaux d'enquêtes.

Cet article consolide l'échantillonnage double en le présentant sous forme unifiée. Plusieurs exemples de sondage utilisés à Statistique Canada illustrent cette unification.

MOTS CLÉS : L'échantillonnage double ; données auxiliaires ; régression ; optimale.

1. INTRODUCTION

La théorie de l'échantillonnage double est habituellement présentée en supposant que l'un des échantillons est imbriqué dans l'autre. Ce genre de sondage est appelé sondage à deux phases. L'échantillon de première phase fournit de l'information auxiliaire (x) relativement peu chère à obtenir, alors que l'échantillon de deuxième phase contient les variables d'intérêt. On se sert des données de la première phase de plusieurs façons : (a) pour stratifier l'échantillon de deuxième phase ; (b) pour améliorer l'estimation en se servant de l'estimateur par différence, quotient ou régression ; (c) pour sous échantillonner un ensemble d'unités non répondantes. L'échantillonnage à deux phases est une technique puissante qui a une longue histoire. Neyman (1938) a été le premier à la proposer. Rao (1973) a étudié l'échantillonnage double dans le contexte de la stratification et des études analytiques. Cochran (1977) présentait les résultats fondamentaux de l'échantillonnage à deux phases, y compris les estimateurs de régression les plus simples pour les plans d'échantillonnage de ce genre. Des travaux plus récents en la matière comprennent ceux de Breidt et Fuller (1993), qui ont mis au point des méthodes d'estimation efficaces sur le plan des calculs pour l'échantillonnage à trois phases, en présence de données auxiliaires. Chaudhuri et Roy (1994) se sont, pour leur part, penchés sur les propriétés optimales des estimateurs de régression plus simples mais bien connus de l'échantillonnage à deux phases. Hidiroglou et Särndal (1998) ont suggéré des estimateurs basés sur le calage et la régression pour l'échantillonnage à deux phases afin de tenir compte de la disponibilité de données auxiliaires aux deux niveaux du plan de sondage.

La théorie de l'échantillonnage double est habituellement présentée en supposant que l'un des échantillons est imbriqué dans l'autre, ou même sélectionné dans la même base de sondage. Ce cas est appelé l'échantillonnage double *non-imbriqué*. Il est légèrement abordé dans les livres classiques du sondage (Des Raj 1968, Cochran 1977). Cette méthode est présentement utilisée dans plusieurs organismes de sondage. L'Enquête canadienne sur l'emploi, la rémunération et les heures (EBRH) menée à Statistique Canada en est un exemple (Rancourt et Hidiroglou 1998). Dans cette enquête, deux échantillons indépendants sont sélectionnés à partir de deux bases de sondage différentes, mais représentent le même univers. Les données auxiliaires (x), qui comprennent le nombre d'employés et sont obtenues d'un échantillon sélectionné d'une base administrative provenant de l'Agence des douanes et du Revenu du Canada. Ces mêmes variables ainsi que les variables d'intérêt (y), le nombre d'heures travaillées par les employés et le sommaire de la rémunération, sont recueillies à partir d'un échantillon sélectionné du Registre des entreprises de Statistique Canada. Deville (1999) décrit une situation semblable à l'INSEE pour une enquête auprès des ménages.

Cependant, ils découlent d'un seul estimateur et ne diffèrent que par leurs variances. Cet article se structure comme suit. La partie 2 expose la notation. La partie 3 présente l'unification de ces deux méthodes de l'échantillonnage double. À la partie 4, on présente la variance estimée, l'estimateur de calage imbriqué et non-imbriqué. On donne plusieurs exemples à la partie 5. La partie 6 récapitule brièvement ce qui a été appris.

¹ M.A. Hidiroglou, Division des méthodes d'enquêtes auprès des entreprises, Édifice R.-H.-Coats, 11-A, Statistique Canada, Ottawa (Ontario), K1A 0T6. Courriel : hidiroglou@statcan.ca

- COWLES, M.K., et CARLIN, B.P. (1996). Markov Chain Monte Carlo convergences diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.
- DEMPSTER, A.P., LAIRD, N.M. et RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (avec discussion). *Journal of the Royal Statistical Society*, B, 39, 1-38.
- DREW, J.H., et FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Alexandria, VA, 623-628.
- ELTINGE, J.L., et YANSANEH, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour les non-réponses à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.
- GELMAN, A., CARLIN, J.B., STERN, H.S. et RUBIN, D.B. (1998). *Bayesian Data Analysis*. Chapitre 14, Generalized Linear Models. London: Chapman & Hall.
- GROVES, R.M., et COOPER, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley and Sons.
- HIEDEBERGER, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de la Statistique*, 54, 139-157.
- LITTLE, R.J.A., et RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley and Sons.
- MACFACHERN, S.N., et BERLINER, L.M. (1994). Subsampling the Gibbs Sampler. *The American Statistician*, 48, 188-189.
- MALLER, R., et ZHOU, X. (1996). *Survival Analysis with Long Term Survivors*. Chichester, UK: Wiley and Sons.
- NATARAJAN, R., et KASS, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227-237.
- NORTHRUP, D.A. (1993). Attitudes Towards Workplace Smoking Legislation: A Survey of Residents of Metropolitan Toronto, Phase III, 1992/93 Documentation techniques. Tech. Rep. Institute for Social Research, York University, non-publiée.
- PEDERSON, L.L., BULL, S.B. et ASHLEY, M.J. (1996). Smoking in the workplace: Do smoking patterns and attitudes reflect the legislative environment? *Tobacco Control*, 5, 39-45.
- PEDERSON, L.L., BULL, S.B., ASHLEY, M.J. et LEFEOE, N.M. (1989). A population survey on legislative measures to restrict smoking in Ontario: 3. Variables related to attitudes of smokers and nonsmokers. *American Journal of Preventive Medicine*, 5, 313-322.
- POTTOFF, R.F., MANTON, K.G. et WOODBURY, M.A. (1993). Correcting for nonavailability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, 88, 1197-1207.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.
- SHMUELL, G., MINKA, T.P., KADANE, J.B., BORLE, S. et BOATWRIGHT, P. (2001). Using Computational and Mathematical Methods to Explore a New Distribution: The V-Poisson. Rapport Technique 740, Department of Statistics Carnegie Mellon University.
- TANNER, M.A. et WONG, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-549.

d'intérêt, cette conjecture est peut-être peu judicieuse. Les indications dont on dispose au moment de la collecte des données peuvent permettre d'évaluer si le mécanisme qui cause la non-réponse est ignorable ou non. On peut donc en conclure qu'il serait bon que les gens qui travaillent avec de telles données utilisent les indications disponibles sur la non-réponse dans leur appréciation de cette information et qu'ils communiquent les indications en question aux autres utilisateurs de l'ensemble de données. En règle générale, la collecte et l'analyse de données qui nous disent où et comment on a trouvé les répondants et combien il a été difficile de les joindre sont pour nous une importante voie qui s'ouvre à la méthodologie d'enquête et à la pratique.

REMERCIEMENTS

Cette étude a été financée par la subvention DMS-9801401 de la National Science Foundation. Les auteurs remercient Shelley Bull de toutes ses observations et ses suggestions utiles et de son aide à l'acquisition des données, tout comme John Eltinge, les critiques anonymes et le rédacteur associé de publication de leurs précieux commentaires.

C'est l'Institute for Social Research de l'Université York qui a fourni les données de l'enquête sur les attitudes à l'égard du règlement sur l'usage du tabac, laquelle a été financée par Santé et Bien-être social Canada. Les données ont été réunies par l'ISR pour le Dr Linda Pederson, de l'Université Western Ontario, et les Drs Shelley Bull et Mary Jane Ashley, de l'Université de Toronto. Les responsables de l'enquête, le ministère ontarien de la Santé et l'Institute for Social Research n'assument aucune responsabilité à l'égard des éléments d'analyse et d'interprétation du présent document.

A. Poststratification

HHW_i est le poids de ménage du sujet i comme il est décrit dans Northrup (1993).

- Soit m le nombre de répondants.
- Soit r le nombre cumulé d'adultes des ménages répondants.
- Soit h_i le nombre d'adultes du ménage du sujet i .
- $HHW_i = h_i \cdot m/r$.

Les proportions de sujets échantillonnés qui appartiennent aux tranches d'âge suivantes ont été calculées pour les répondants des deux sexes : 18-24, 25-44, 45-64 et 65 ans et plus. On a ensuite comparé les pourcentages à la structure par âge-sexe de la population de la région métropolitaine de Toronto.

BIBLIOGRAPHIE

- BIEMER, P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17, 2, 295-320.
- BULL, S. (1994). *Case Studies in Biometry*. Analysis of Attitudes toward Workplace Smoking Restrictions, chapitre 16, New York: Wiley and Sons, 249-270.

La simulation complète MCCM pour le modèle IN consiste en l'application d'un algorithme Metropolis avec pour complément des éléments d'augmentation de données décrits à la section 5.3. Voici un aperçu de cet algorithme. Les variables employées sont définies à la section 5. À chaque itération i ,

1. On tire p_i pour $Beta(s_{i-1} + 1, 2398 - s_{i-1} + 1)$.
2. On impute s_i à partir de $Binomial(p_i) \geq 1,429$.
3. On impute C_{mis_i} : on tire les $(s_i - 1, 429)$ v_i de $Geometric(\pi_{i-1})$ et de $Vc_i \in C_{mis_i}$, $c_i = v_i + 12$.
4. On tire π_i de $Beta(s_i + 1, \sum C_{sus_i} - s_i + 1)$.
5. On impute les valeurs du reste de X_{mis_i} en utilisant les relations avec le nombre d'appels comme le décrit la section 5.3.
6. On met à jour les paramètres supplémentaires utilisés dans l'augmentation de données pour X_{mis_i} .

- On met à jour les paramètres de régression linéaire β et σ^2 en se reportant directement à la forme fermée de leurs distributions postérieures.
- On met à jour les paramètres de régression logistique β_i par application par pas de l'algorithme Metropolis à chacun.
- 7. On impute Y_{mis_i} : $\forall y_i \in Y_{mis_i}$; on tire y_i de $Multinomial(p_0(x_i), p_1(x_i), p_2(x_i))$.
- 8. On met à jour chaque β_{kj} par application d'une itération de l'algorithme Metropolis sur la vraisemblance conditionnelle et par application d'une fonction de saut normale.

B. Simulation MCCM

- Soit p_{1i} la proportion des résidents d'âge adulte de la région métropolitaine de Toronto qui appartiennent à la catégorie d'âge-sexe du sujet i selon les données du recensement de 1991.
- Soit p_{2i} la proportion des répondants appartenant aux catégories d'âge et de sexe du sujet i .
- $W_i = HHW_i \cdot p_{1i}/p_{2i}$, où W_i est le poids final de poststratification utilisé dans l'analyse.

Tableau 2
Comparaison des probabilités de réponse pour quatre sujets types. Nous avons utilisé les médianes postérieures comme estimations ponctuelles des coefficients des modèles bayésiens; nous avons employé la valeur EMV pour le modèle fréquentiste

Sujet 1	Sujet 2	Sujet 3	Sujet 4
État de fumeur Non	Non	Ancien	Oui
Âge 30	50	27	40
État de réaction D'habitude	Toujours	Non	Non
Connaissance des risques 11	12	7	3
$Y=1/Y=0$			
Probabilités du modèle	2,105	0,457	0,396
MAR EMV	0,674		
MAR distribution antérieure en valeurs diffusées	0,703	4,487	0,009
0,116			
IN distribution antérieure en valeurs diffusées : 2 appels	0,640	4,024	0,202
0,108			
IN distribution antérieure en valeurs centrales : 2 appels	0,593	4,442	0,162
0,102			
Option 3 : 2 appels	0,594	4,449	0,162
0,102			
Option 4 : 2 appels	0,592	4,435	0,162
0,101			
Option 5 : 2 appels	0,590	4,423	0,161
0,101			
Option 6 : 2 appels	0,590	4,426	0,161
0,101			
IN distribution antérieure en valeurs diffusées : 13 appels	0,974	6,128	0,308
0,165			
IN distribution antérieure en valeurs centrales : 13 appels	0,936	7,013	0,256
0,160			
Option 3 : 13 appels	0,937	7,026	0,256
0,161			
Option 4 : 13 appels	0,934	7,000	0,255
0,160			
Option 5 : 13 appels	0,930	6,975	0,254
0,159			
Option 6 : 13 appels	0,931	6,980	0,254
0,160			

7.4 Effet sur les probabilités de réponse

La variation des probabilités postérieures que nous venons de décrire s'accompagne d'une variation des probabilités de réponse estimées d'un sujet dans une catégorie. Parmi les répondants, 57,45 % ont choisi la catégorie 0, 40,64 %, la catégorie 1, et 1,91 %, la catégorie 2. Pour le nombre de non-répondants non réfractaires, on relève une médiane postérieure de 469 et un intervalle crédible à 95 % de (25, 944). En moyenne, 55,88 % des non-répondants non réfractaires simulés ont choisi la catégorie 0, 40,03 %, la catégorie 1, et 4,08 %, la catégorie 2. Bien que, pour les catégories 0 et 1, les valeurs moyennes des non-répondants non réfractaires tombent bel et bien dans les intervalles de confiance à 95 % pour les proportions de répondants dans ces catégories, les estimations ponctuelles varient pour chaque catégorie en cas d'inclusion du mécanisme de non-réponse dans le modèle. Dans une comparaison des résultats de la catégorie 2, nous estimons que les non-répondants ont deux fois plus de chances que les répondants de choisir « Permissio

non-répondants dans cette catégorie illustre comment on peut tirer de fausses conclusions au sujet des données si on ne tient pas bien compte des non-répondants.

8. CONCLUSION

La section 7 démontre que, pour la variable dépendante d'intérêt dans cet ensemble de données, l'affirmation que les observations manquantes sont aléatoires avant prise en compte du mécanisme de non-réponse se révèle erronée. Ceci suppose que la relation entre les variables pertinentes est la même pour tous les sujets non réfractaires. Ajoutons que l'application d'une fausse hypothèse MAR dans l'évaluation de cette variable dépendante risque d'entacher d'une grave erreur le calcul des probabilités postérieures et toute conclusion à en tirer. Pour bien évaluer les options sur l'usage du tabac en milieu de travail à Toronto au début de 1993 par la variable dépendante d'intérêt dans le cadre de cette enquête, il est nécessaire d'intégrer le mécanisme de non-réponse à la structure du modèle.

Dans notre analyse, nous avons utilisé un seul élément d'information, le nombre d'appels. Le traitement aurait pu être plus complet si nous avions disposé de plus de renseignements. Si nous avions connu le nombre exact de tentatives d'entrée en communication avec les non-répondants – au lieu d'un minimum d'appels –, ainsi que l'heure des appels, l'analyse aurait gagné en précision. Qui plus est, si nous avions connu la nature de la non-réponse par refus ou indisponnibilité et le nombre effectif de tentatives d'entrée en communication avec les non-répondants, il aurait été possible de mieux caractériser ces derniers. Groves et Couper (1998) signalent que les erreurs statistiques différeront probablement selon qu'il s'agit d'une non-réponse par indisponnibilité ou par refus. Comme ils le précisent, un important point en recherche est l'évaluation de l'incidence pour contacter les enquêtés sur l'erreur de mesure.

Les résultats que nous avons présentés ne valent que pour cette variable dépendante évaluant l'usage du tabac en milieu de travail dans ce seul ensemble de données. Comme on peut percevoir que le tabagisme est devenu moins socialement acceptable ces dernières années, on serait fondé à penser que l'erreur de non-réponse due aux questions sur l'usage du tabac pourrait être plus sérieuse que pour d'autres questions. On peut trouver dans Biemer (2001) une comparaison de biais de non-réponse pour diverses questions sur le tabagisme et d'autres sujets. La comparaison n'accrédite pas l'idée que l'erreur de non-réponse est propre aux questions portant sur l'usage du tabac.

De nos résultats, il n'y a rien à tirer comme implications au sujet des mécanismes de non-réponse d'autres enquêtes, mais on peut clairement voir ici que, si on suppose – à tort – que les répondants d'une enquête constituent un sous-échantillon aléatoire d'une population pour les variables

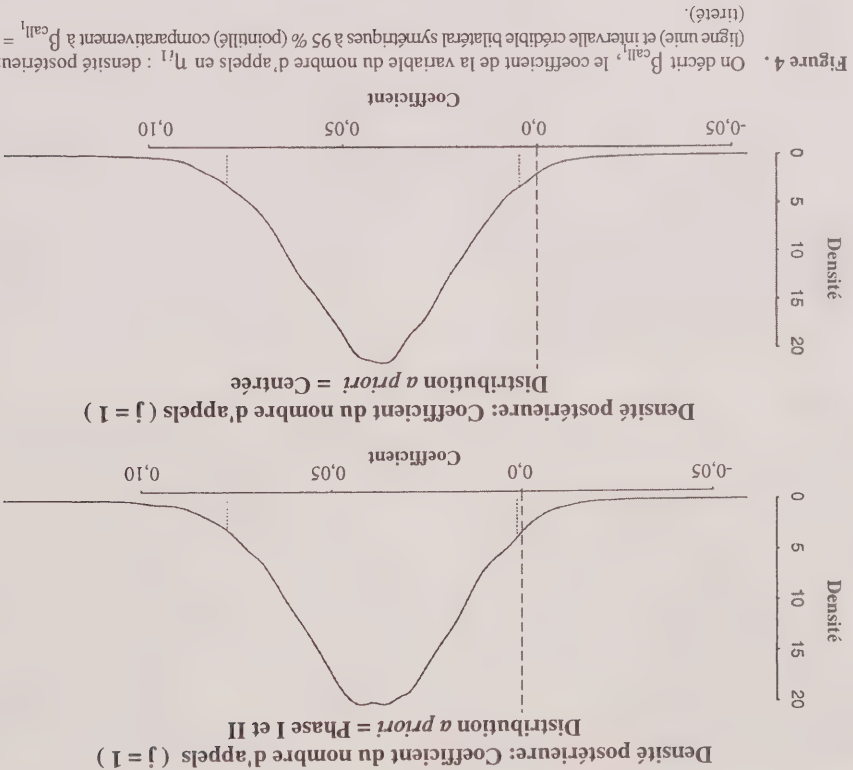


Figure 4.

On décrit β_{call} , le coefficient de la variable du nombre d'appels en M_{11} : densité postérieure (ligne unie) et intervalle crédible bilatéral symétriques à 95 % (pointillés) comparativement à $\beta_{call} = 0$ (tireté).

7.3 Effet sur les probabilités de réponse

L'hypothèse MAR étant jugée lacunaire, il est bon de s'interroger sur l'utilité de l'erreur que créerait son application. Il est possible d'illustrer l'ordre de grandeur de l'erreur consécutive à l'adoption d'une fausse hypothèse MAR en s'attachant à son effet sur le rapport des probabilités $p_1(x_1)/p_0(x_1)$. Considérons d'abord l'effet sur un profil type de répondant. Le répondant modal est un non-fumeur de 25 à 35 ans que dérange d'habitude la fumée secondaire, qui présente une valeur « Connaissance des risques » de 11 et qui est joignable en deux appels. Nous faisons de ce répondant modal notre sujet 1. Le tableau 2 montre le changement de probabilités postérieures de ce sujet s'il est appelé 13 fois.

La colonne « Sujet 1 » du tableau 2 fait voir une différence considérable de probabilités postérieures si on tient compte du mécanisme de non-réponse. Pour ce profil type de répondant, si le nombre d'appels monte de 2 à 13, les probabilités postérieures de choix de « Interdiction totale de fumer » plutôt que de « Permission de fumer seulement dans des zones réservées » s'accroissent de 52,18 % dans une distribution diffuse et de 57,84 % dans une distribution centrée. C'est la preuve éloquente de l'existence d'un lien entre la variable dépendante et le mécanisme de non-réponse.

S'agit-il de résultats propres à notre sujet modal? Le tableau 2 décrit aussi les effets sur les probabilités de réponse dans le modèle NI pour trois autres profils de sujet en fonction de nos six distributions *a priori*. Le sujet 2 est un non-fumeur de 50 ans que dérange toujours la fumée et qui a une note parfaite « Connaissance des risques ». Le sujet 3 est un ex-fumeur de 27 ans que la fumée ne dérange pas et qui obtient une note de 7 à « Connaissance des risques ». Le sujet 4 est un fumeur de 40 ans que ne dérange pas la fumée et qui reçoit, lui, une note de 3 pour cette même connaissance des risques. Pour des sujets et des distributions multiples, ce tableau dégage invariablement le même résultat. Si on porte le nombre d'appels à plus de 12, on se trouve à augmenter les probabilités postérieures de choix de la catégorie 1 par rapport à la catégorie 0. Pour les divers sujets et distributions du tableau 2, le taux d'accroissement se situe entre 52,18 % et 58,41 %.

On obtient des résultats semblables lorsqu'on examine les probabilités de choix de la catégorie « Permission totale de fumer » par rapport à la catégorie « Permission de fumer seulement dans des zones réservées ». Avec des sujets qui sont un fumeur et un ex-fumeur (sujets 3 et 4), les probabilités postérieures s'élèvent de 46,7 % si on porte le nombre d'appels de 2 à 13 dans une distribution diffuse.

La figure 3 semble indiquer que les probabilités d'appel fructueux décroissent à mesure que s'élève le nombre d'appels. Pour vérifier l'hypothèse de l'existence d'une relation linéaire entre le nombre d'appels et les probabilités de réponse en expression logarithmique, nous avons élaboré un autre modèle bayésien NI qui scinde la variable du nombre d'appels en deux, $C_{(C,27)}^{I(1)}$ et $C_{(C,27)}^{I(2)}$, selon que le nombre d'appels est de moins de sept ou non. Nous avons ensuite comparé les distributions postérieures des coefficients de ces deux variables sans découvrir la preuve qu'elles étaient essentiellement différentes. Précisons que, pour η_{11} , l'intervalle crédible à 95 % pour $C_{(C,27)}^{I(1)}$ contenait le même intervalle pour $C_{(C,27)}^{I(2)}$ et que, pour η_{22} , les intervalles correspondants étaient largement en chevauchement.

7.2 Sensibilité aux distributions a priori

Des distributions a priori différentes pour le coefficient du nombre d'appels ou les autres influencerait-elles sur l'effet que nous avons indiqué? Le tableau 1 présente les intervalles crédibles DPH à 95 % pour le coefficient de la variable du nombre d'appels dans la première équation logistique du modèle NI, et ce, pour six distributions a priori différentes : distributions diffusées et centrées et quatre autres appelées options 3, 4, 5 et 6. Les options 3 et 4 ressemblent à la distribution centrées sauf que la distribution a priori change respectivement à normale (1,9) et à normale (-1,9) pour le coefficient du nombre d'appels. L'option 5 utilise des distributions a priori normales (0,9) pour β_{call} , β_{age} , et $\beta_{b_{USUL}}$, (1,9) pour β_{01} , et (-5,9) pour β_{R-risk} , (-1,9) pour β_{S_1} et $\beta_{b_{NO}}$. Quant à l'option 6, elle prend la distribution a priori centrées et réduit toutes les variances de 9 à 2.

Le tableau 1 démontre que, dans les six options, le coefficient de la variable du nombre d'appels diffère nettement de zéro dans la première équation logistique. Le choix d'une distribution a priori parmi les six options ne semble pas influencer sur la constatation que le mécanisme de non-réponse est non-ignorable pour cet ensemble de données.

Tableau 1

Intervalles crédibles DPH à 95 % pour β_{call} dans six distributions a priori		
Distribution a priori	Coefficient du nombre d'appels	« $C_{I,27}$ » en η_{11}
Intervalles à 95 %		
Borne inférieure	Borne supérieure	
Phase I & II	0,00129	0,07746
Centrée	0,00446	0,07980
Option 3	0,00447	0,07983
Option 4	0,00441	0,07975
Option 5	0,00440	0,07970
Option 6	0,00436	0,07944

À notre avis, une variance de neuf convient sans être trop diffuse. L'utilisation d'une distribution a priori impropre pourrait donner une simulation de Monte Carlo à chaîne de Markov sans convergence possible. De plus, comme l'indiquent Nataraajan et Kass (2000), une distribution propre qui est excessivement diffuse peut se comporter comme une distribution impropre. À la section 7.2, nous évaluons par analyse de sensibilité comment le choix d'une distribution a priori influe sur les résultats.

Les paramètres de non-réponse du modèle NI, π et π_1 , ont eu le même traitement dans l'une et l'autre de ces possibilités. Nous ne disposons pas d'indications supplémentaires sur les probabilités « sujet joint » ou « sujet non réfractaire ». Ainsi, π et π_1 se sont chacun vu attribuer une distribution a priori $U(0,1)$.

Les paramètres d'augmentation de données de chacune des équations de régression logistique β_j ont reçu indépendamment une distribution a priori diffuse (0,9). Pour chaque équation de régression linéaire dans la procédure d'augmentation de données, les coefficients β_j et la variance σ_j^2 ont été fixés à $p(\beta_j, \sigma_j^2) \propto 1/\sigma_j^2$, c'est-à-dire à la distribution a priori non informative type (Gelman et coll. (1998), par exemple). À noter que les formes fermées des distributions postérieures des paramètres de régression linéaire sont connues et peuvent directement s'obtenir.

7. RÉSULTATS

Nous examinons d'abord le bien-fondé de l'hypothèse MAR par les coefficients de la variable du nombre d'appels. Nous évaluons ensuite le modèle NI dans sa sensibilité au choix d'une distribution a priori. Nous étudions enfin l'ordre de grandeur des effets d'une fausse hypothèse MAR pour cet ensemble de données en présentant les changements de probabilités de réponse.

7.1 Coefficients de la variable du nombre d'appels

Pour les distributions a priori tant diffuse que centrées, la figure 4 décrit la densité postérieure (ligne unie) et les intervalles crédibles estimatifs à 95 % (pointillés) du coefficient de la variable du nombre d'appels en η_{11} dans le modèle NI et compare les valeurs au point $\beta_{call} = 0$ (tirets). Les résultats indiquent clairement que ce coefficient est différent de zéro. Nous relevons aussi un résultat non nul en η_{12} où, par la distribution a priori diffuse, l'intervalle crédible DPH à 95 % pour β_{call} est (-0,03613, 0,11595). Le coefficient non nul de $C_{I,27}^{I(2)}$ démontre une dépendance entre le nombre d'appels et l'opinion du sujet sur l'usage du tabac en milieu de travail. Ainsi, la variable dépendante et le mécanisme de non-réponse ne sont pas indépendants dans les conditions dont parle la section 5.2, d'où l'implication d'une fausseté pour cet ensemble de données de l'hypothèse du caractère aléatoire des observations manquantes avant prise en compte du mécanisme de non-réponse.

6. CHOIX DE DISTRIBUTIONS A PRIORI

Dans l'évaluation de distributions *a priori* possibles pour les paramètres des modèles NI et MAR, nous avons tenu compte de l'objectif de comparaison des divers modèles. Le choix de distributions *a priori* pour les paramètres s'est opéré d'un point de vue MAR. Deux possibilités ont été étudiées.

La première s'articule autour de l'exploitation des données des enquêtes des phases I et II. Comme ces enquêtes ont évidemment précédé la phase III (d'où viennent nos données) où l'enquête a été identique, nous pouvons élaborer des distributions *a priori* à partir des données des deux premières phases. On y trouve la même variable dépendante, ainsi que les variables « Etat de fumeur », « Âge » et « Connaissance des risques ». C'est de ces données que nous avons tiré un modèle de régression logistique afin de décrire le rapport entre les options au sujet de l'usage du tabac en milieu de travail et ces trois variables explicatives. Nous avons établi des distributions antérieures normales pour les coefficients des trois à leurs valeurs centrales EMV, mais avec une erreur-type majorée. Nous avons accru les termes d'erreur pour trois raisons :

(i) trois ans s'étaient écoulés entre la phase II et la phase III et les opinions auraient pu changer dans ce laps de temps à cause de l'incidence du règlement municipal;

(ii) les valeurs EMV ont été calculées à l'aide de l'hypothèse MAR même qu'il était évaluée;

(iii) avant la collecte des données de phase III, il était possible que d'autres variables explicatives figurent dans le modèle et, du fait de leur présence, l'effet des trois variables considérées pourrait être autre.

Les variances ont augmenté, mais les moyennes sont restées les mêmes, car on ignorait au départ quel serait le sens de toute variation. Comme les données disponibles des phases I et II ne renseignaient pas sur le nombre d'appels ni sur l'état de réaction, on a attribué aux coefficients de ces variables une distribution *a priori* normale diffuse (0,9). Pour plus de clarté, nous appellerons cette première possibilité « distribution *a priori* des phases I et II » dans notre analyse.

La seconde possibilité consiste à attribuer une distribution *a priori* normale (0,9) aux divers coefficients de régression logistique. Ce choix repose sur les trois raisons mêmes pour lesquelles nous avons accru les termes d'erreur plus haut, c'est-à-dire parce que les variables communes aux enquêtes des phases I et II et de la phase III ne sont pas échangeables. Une élaboration fondée sur les résultats des phases I et II serait peu appropriée. Nous appelons cette seconde possibilité « distribution *a priori* centrée ».

Si nous optons ici pour une distribution normale (0,9), c'est par commodité. Si on centre la distribution *a priori* à zéro, on prête un même poids à l'un et l'autre sens de la

(2 398, p), où $1.429 \leq S \leq 2.398$. Étant donné S , le nombre de sujets dans A_{mis} est connu. Pour chacun de ces sujets, on peut tirer une valeur $V_i \sim \text{Geometric}(\pi)$, d'où une imputation pour le nombre d'appels dont on a besoin pour joindre chaque sujet non réfractaire et non joint. On peut alors exploiter les relations entre le nombre d'appels et les autres variables explicatives afin d'imputer des valeurs pour le reste de X_{mis} . Plus précisément, on procède à l'imputation des valeurs manquantes « Âge » et « Connaissance des risques » en effectuant respectivement des régressions du nombre d'appels sur ces variables et en dégageant une prévision des équations linéaires ainsi obtenues. De même, les valeurs manquantes de « Etat de fumeur » et de « Etat de réaction » s'imputent par régression logistique dans l'un et l'autre cas avec le nombre d'appels comme variable explicative. Ici, on vérifie les hypothèses du modèle à l'aide des données relatives aux répondants et pose l'hypothèse que les mêmes relations valent pour les non-répondants non réfractaires. Il convient de noter que ces équations de régression ordinaire et de régression logistique s'insèrent dans un contexte bayésien (Gelman, Carlin, Stern et Rubin 1998) et qu'il faut inclure d'autres paramètres, β_j , dans la le processus MCMC qui décrit ces relations (on trouvera plus de détails à l'annexe B). Si nous choisissons ce plan d'imputation, c'est par souci d'efficacité de tout l'algorithme MCMC. Comme solution de rechange, il y aurait l'imputation des valeurs manquantes d'une variable explicative en particulier en conditionnant par toutes les autres variables (Rubin 1996, par exemple). Enfin, Y_{mis} peut faire l'objet d'une prévision par les valeurs d'imputation de X_{mis} et la relation décrite dans le modèle de régression logistique. Par souci d'échangeabilité des non-répondants non réfractaires et faute de données de poststratification, nous appliquons une valeur de pondération de 1,0 à toutes les valeurs imputées Y_{mis} . Comme autre choix, nous pouvons – en dehors de l'âge – imputer le sexe et la taille du ménage pour les non-répondants non réfractaires et appliquer la méthode de pondération que décrit l'annexe A aux valeurs imputées Y_{mis} .

5.4 Échantillonnage de distribution postérieure

Tout l'exercice de simulation MCMC consiste en l'application d'un algorithme Metropolis avec enrichissement à chaque itération par la technique d'augmentation de données que nous venons de décrire. L'algorithme MCMC utilisé est exposé sommairement à l'annexe B. Nous évaluons la convergence par la méthode de Hiedelberger et Welch (1983) décrite dans Cowles et Carlin (1996). MacEachern et Berliner (1994) affirment que, dans des conditions non strictes, le sous-échantillonnage des valeurs de simulation MCMC en fonction de l'autocorrélation donnera des estimateurs moins bons. Voilà pourquoi nous avons utilisé toutes les valeurs de simulation dans l'analyse après la période nécessaire d'itérations.

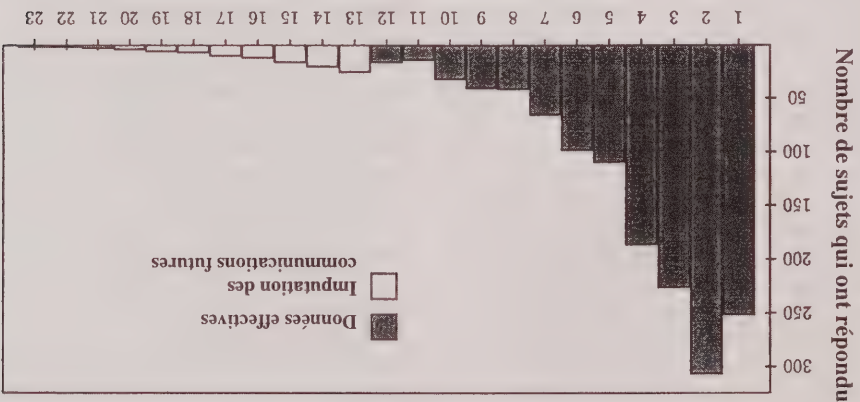


Figure 3. On décrit le nombre effectif d'entrées en communication pour chacune des 12 premières tentatives, et le nombre prévu d'entrées en communication aux treizième et suivantes. Les valeurs d'imputation sont fondées sur des probabilités de tentative fructueuse de 0,205 et de « sujet non réfractaire » de 0,636.

base du modèle selon laquelle la relation entre le nombre d'appels, la variable dépendante et les autres variables explicatives considérées est la même pour les répondants et les non-répondants. Si on prend le nombre d'appels en compte au volé « régression logistique » du modèle, on se trouve à exclure les sujets épargnés, car il n'y aura jamais de contact avec eux.

La fonction de pseudovraisemblance totale du modèle IN (ou plus précisément du modèle IN des sujets non réfractaires) est le produit des morceaux « non-réponse » et

« régression logistique » :

$$L(p, \pi, \beta) \propto \left[p^m \pi^m (1 - \pi)^{\left(\sum_{i=1}^m c_i \right) - m} \right] \times \left[(1 - p) + p (1 - \pi)^{12 - m} \right] \times \left[\prod_{j=1}^m \prod_{i=1}^2 \frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right]^{y_{ij} w_i} \quad (6)$$

À noter que nous incluons ici la variable W_i de pondération de ménage et de poststratification en vue de tenir compte de ce qu'une stratification appropriée des répondants pourrait rendre inutile l'introduction d'un mécanisme de description de la non-réponse.

5.3 Augmentation de données

Tanner et Wong (1987) proposent un calcul par itération augmentant l'ensemble de données, l'analyse devient plus simple et que les éléments d'augmentation sont facilement

produits. Il faut compléter la notation. Soit S le nombre total de sujets non réfractaires dans l'échantillon. $S = \sum_{i=1}^n Z_i$, $S \sim \text{Binomial}(p)$. Soit X la matrice des variables explicatives (avec le nombre d'appels) pour l'ensemble des sujets sélectionnés aux fins de l'enquête. Soit $Y = (Y_1, \dots, Y_n)$ le vecteur de leurs réponses. On divise X en $\{X_{\text{obs}}^{\text{mis}}, X_{\text{imm}}^{\text{mis}}, X_{\text{imm}}^{\text{mis}}\}$ et Y en $\{Y_{\text{obs}}^{\text{mis}}, Y_{\text{imm}}^{\text{mis}}, Y_{\text{imm}}^{\text{mis}}\}$. On sait, par la propriété sans mémoire de la distribution géométrique, quelle est la distribution du nombre supplémentaire d'appels dont on a besoin pour joindre les sujets dans A_{mis} . On peut ainsi l'exprimer : $V_i \in A_{\text{mis}}$, soit $V_i = C_i - 12$, distribuée selon la loi géométrique de paramètre π .

Supposons maintenant que les valeurs vraies de S , X_{mis} et Y_{mis} sont connues. Le calcul de vraisemblance pourrait alors prendre la forme suivante :

$$L(p, \pi, \beta | X_{\text{obs}}^{\text{mis}}, X_{\text{imm}}^{\text{mis}}, Y_{\text{obs}}^{\text{mis}}, Y_{\text{imm}}^{\text{mis}}, S, R) \propto \left[(p\pi)^s (1 - \pi)^{\left(\sum_{i=1}^s c_i \right) - s} \right] \times \left[(1 - p)^{n-s} \right] \times \left[\prod_{j=1}^s \prod_{i=1}^2 \frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right]^{y_{ij} w_i} \quad (7)$$

On ignore les valeurs vraies de S , X_{mis} et Y_{mis} , mais on approprié de sujets. réfractaires et où les sommes se prennent sur l'ensemble dont on aurait eu besoin pour joindre tous les sujets non d'appels où $\sum C_{\text{sus}}^{\text{sus}} = \sum C_{\text{obs}}^{\text{sus}} + \sum (V_i + 12)$ est le nombre d'appels

imputer des valeurs stochastiquement possibles dans l'algorithme MCMC. Étant donné p , on peut tirer une valeur de S d'une distribution binomiale tronquée

en communication avec les sujets non réfractaires suit une distribution géométrique, c'est-à-dire que $C_i^1/Z_i^1 = 1 \sim \text{Geometric}(\pi)$. Se trouve-t-on à éliminer le problème illustré à la figure 2?

Soit R_i un indicateur de réponse du sujet i . On peut prendre le mécanisme de non-réponse en compte par l'intégration de ces indicateurs de réponse au modèle. Il reste que l'introduction de la variable « sujet non réfractaire » implique deux catégories distinctes de non-réponse. Il est donc possible de faire une caractérisation plus fine et d'utiliser tant les indicateurs de sujet non réfractaire $Z = (Z_1^1, \dots, Z_n^1)^T$ et de réponse R dans un modèle mixte de description de non-réponse. Dans une mise à jour de l'équation (1), le mécanisme de non-réponse est non informatif si et si seulement (π, p) est distinct de θ et que

$$f(R, Z | Y^{\text{obs}}, Y^{\text{mis}}, \pi, p) = f(R, Z | Y^{\text{obs}}, \pi, p). \quad (4)$$

Soit $C^{\text{obs}} = (C_1^1, \dots, C_m^1)$ et $Z^{\text{obs}} = (Z_1^1, \dots, Z_m^1)$ les vecteurs du nombre d'appels et du caractère « non réfractaire » observé de chaque enquête. Soit $R = (R_1^1, \dots, R_n^1) = \text{le vecteur de réponse pour chaque sujet visé. Chacun des sujets } i \text{ peut se ranger par sa réponse dans trois catégories qui s'excluent les unes les autres, à savoir } A^{\text{obs}} - \text{observé, } A^{\text{mis}} - \text{manquant et } A^{\text{imms}} - \text{épargné, où :}$

$$A^{\text{obs}} = \{i : i \text{ était non réfractaire et a répondu}\}$$

$$A^{\text{mis}} = \{i : i \text{ était non réfractaire, mais n'a pas répondu}$$

$$\text{en 12 tentatives d'entrée en communication}\}$$

$$A^{\text{imms}} = \{i : i \text{ était réfractaire}\}.$$

Les probabilités d'inclusion d'un sujet dans chacune de ces catégories peuvent ainsi se calculer :

$$P(i \in A^{\text{obs}}) = P(Z_i^1 = 1, R_i^1 = 1, C_i^1 = c_i^1) = p\pi(1 - \pi)^{c_i^1 - 1}$$

$$P(i \in A^{\text{mis}}) = P(Z_i^1 = 1, R_i^1 = 0, C_i^1 > 12) = p(1 - \pi)^{12}$$

$$P(i \in A^{\text{imms}}) = P(Z_i^1 = 0) = 1 - p.$$

Selon ces données, $m = 1\,429$ sujets dans A^{obs} et $n - m = 969$ sujets non répondants dans $A^{\text{mis}} \cup A^{\text{imms}}$; $n = 2\,398$ est le nombre estimatif total de sujets sélectionnés. Ainsi, la densité conjointe de Z^{obs} , R et C^{obs} étant donné p et π est la suivante :

$$f(Z^{\text{obs}}, R, C^{\text{obs}} | p, \pi) \propto$$

$$\left[p^m \pi^m (1 - \pi)^{\sum_{i=1}^m c_i^1 - m} \times \left[(1 - p) + p(1 - \pi)^{12} \right]^{n-m} \right]. \quad (5)$$

Le modèle mixte décrit par l'équation 5 peut être considéré comme un cas spécial des modèles de non-réponse examinés dans Drew et Fuller (1981).

Il serait bon de vérifier si cette distribution conjointe est une juste représentation des tendances de réponse des sujets « non réfractaires » dans l'ensemble de données. L'estimation EMV de p est simplement la proportion de répondants dans l'échantillon, et on se trouve à nettement sous-estimer p . Si on pose les distributions *a priori* $U(0, 1)$ pour p et π à la fois et examine leur distribution conjointe postérieure par simulation MCMC, les médianes postérieures s'établissent à $p = 0,636$ et à $\pi = 0,205$, avec des intervalles postérieurs crédibles bilatéraux symétriques de (0,613, 0,659) et (0,191, 0,219) pour p et π respectivement. La figure 3 indique à quoi ressemblerait l'ensemble de données après imputation du nombre manquant d'appels pour les non-répondants non réfractaires selon les médianes postérieures. On se trouve à avoir éliminé en majeure partie le problème de la figure 2.

La distribution géométrique paraît suffisante (après prise en compte du caractère non réfractaire), mais un critère s'est interrogé sur l'utilisation de cette distribution sans prise en compte de covariables peut-être utiles. Comme nous l'avons expliqué, on n'a pas recueilli de données pour les covariables qui, selon nous, auraient été des plus intéressantes aux fins de cette analyse. Une autre possibilité de modélisation du mécanisme de réponse des sujets non réfractaires est l'emploi d'une distribution Gamma discrète-tisser. Si on a besoin de plus de complexité, on peut aussi songer à la distribution V-Poisson (une distribution Poisson biparamétrique qui généralise un certain nombre de distributions discrètes bien connues, dont la distribution géométrique) de Shmueli, Minka, Kadane, Borle et Boatwright (2001).

5.2 Rattachement de la non-réponse à la variable dépendante – modèle NI

Comme la distribution géométrique conditionnelle du nombre d'appels décrit la non-réponse des sujets « non réfractaires », on peut tenir compte de l'effet de cette non-réponse sur la variable dépendante en faisant du nombre d'appels une variable explicative supplémentaire aux fins de l'estimation de vraisemblance en régression logistique. On se trouve ainsi à ajouter deux paramètres au volet « régression logistique » du modèle. Ce sont les coefficients β_{call} de chacune des équations linéaires η_{ij} décrites à l'équation (2).

L'existence de coefficients non nuls du nombre d'appels indiquerait alors que la variable dépendante n'est pas indépendante du mécanisme de non-réponse et que, par conséquent, celui-ci n'est pas ignorable. Si les coefficients sont nuls, la non-réponse des « non réfractaires » est ignorable. Les conclusions tirées ici s'appuient sur l'hypothèse à la

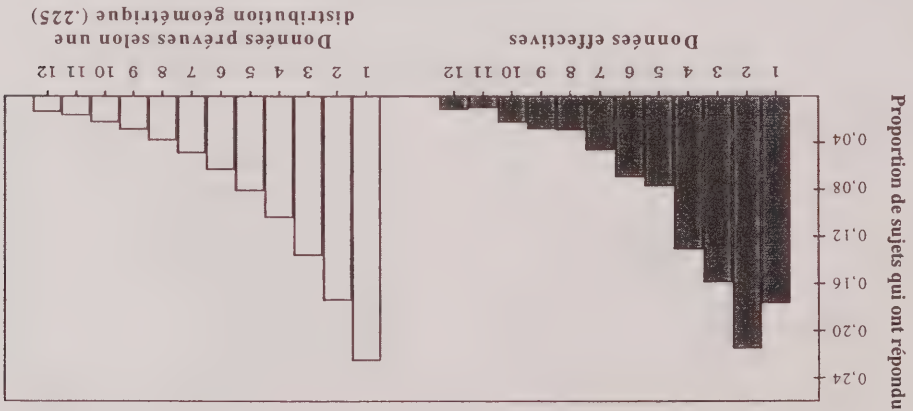


Figure 1. On compare les données effectives d'enquête sur les entrées en communication aux 12 premières tentatives, d'une part, et les résultats prévus selon une distribution géométrique (paramètre $\pi : 0,225$) du nombre d'appels nécessaires à l'achèvement de l'interview, d'autre part.

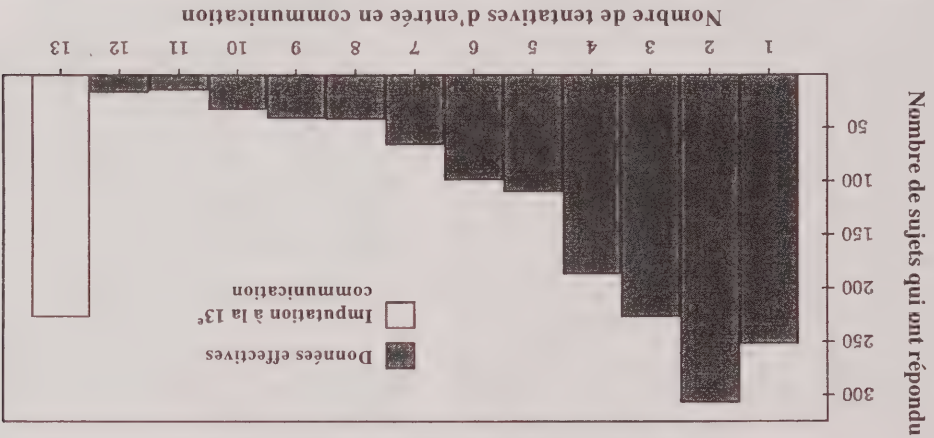


Figure 2. On décrit le nombre effectif d'entrées en communication pour chacune des 12 premières tentatives, ainsi que le nombre prévu d'entrées en communication à la 13e tentative. La valeur probable pour la 13e tentative est tirée d'une distribution géométrique ($\pi : 0,225$) où on modélise le nombre d'appels jusqu'à l'achèvement de l'interview.

Compte tenu de ces données, il est possible d'affirmer que ce ne sont pas tous les sujets sélectionnés qui seront joignables. Maillet et Zhou (1996) décrivent les sujets éparpillés, c'est-à-dire ceux qui ne subissent pas l'événement d'intérêt. Pour reprendre les termes qu'ils emploient, s'il est impossible d'obtenir une réponse d'un sujet sélectionné après un nombre illimité d'appels, le sujet est caractérisé comme « éparpillé ». Les sujets « non éparpillés » sont alors caractérisés comme « non éparpillés ». L'ensemble des sujets éparpillés (c'est-à-dire réfractaires) forme alors la fraction de « survivants » de l'échantillon. Pour employer des termes plus familiers, les sujets éparpillés sont ceux qui, joints, ont refusé de répondre, qui auraient refusé s'ils avaient été joints ou qui étaient physiquement

ou mentalement incapables de jamais participer. Northrup (1993) indique que ceux qui ont initialement refusé de participer ont ensuite été joints par les intervieweurs les plus expérimentés, aussi posons-nous l'hypothèse ici que tous ceux qui ont continué à refuser n'auraient jamais participé. Le groupe des non-réfractaires comprend les enquêtés qui, joints, auraient répondu et ceux qui étaient physiquement ou mentalement incapables de participer à l'époque de la collecte de données, mais qui auraient été désireux et capables de le faire en tout autre temps. Soit la variable $Z_i = I_i^{(\text{non réfractaire})}$ (sujet i) comme indicateur de sujet i non réfractaire et $p = P(\text{sujet } i \text{ non réfractaire})$, c'est-à-dire $Z_i \sim \text{Bernoulli}(p)$. Supposons maintenant que le nombre de tentatives d'entrée

5. MODÈLE NI

5.1 Modélisation du mécanisme de non-réponse

Comme les valeurs manquantes ne sont pas nécessairement aléatoires, nous devons tenir compte du mécanisme qui cause la non-réponse. Northrup (1993) indique que les non-répondants à l'enquête ont reçu au moins 12 appels infructueux (dont 3 le jour, 4 le soir et 4 le week-end au moins). Nous devons malheureusement constater que d'autres indications utiles du nombre d'appels n'ont pas été retenues. Nous ignorons lesquels des non-répondants ont reçu plus de 12 appels ou quelles ont été les périodes d'appel (jour, soir ou fin de semaine). Nous ignorons également les détails de la non-réponse, c'est-à-dire si le sujet a refusé de participer après entrée en communication, s'il y a jamais eu enregistrement du message d'appel par un répondant ou si on n'a pas répondu du tout. La stratification des non-répondants est donc impossible et ceux-ci sont tous considérés comme unités interchangeables dans la présente analyse.

On a tenu un certain nombre de fois d'entrer en communication avec chaque sujet jusqu'à achievement d'interview ou caractérisation de non-répondant. Dans le cas des répondants, la variable du nombre d'appels (C_j^i) indique le nombre de tentatives infructueuses jusqu'à la première entrée en communication. Nous pouvons nous attendre, par conséquent, à ce que le nombre d'appels suive une distribution géométrique avec troncation des observations pour les non-répondants. Plus précisément, soit $\pi = P$ (tentative fructueuse). Considérons alors $C_j^i \sim \text{Geometric}(\pi)$ et $P(C_j^i = c_j^i) = \pi(1 - \pi)^{c_j^i - 1}$. À noter que, si nous avions disposé de données auxiliaires sur le nombre d'appels dans le cas des non-répondants (comme dans Groves et Couper 1998), nous pourrions avoir calculé ici des probabilités conditionnelles de réponse.

Les histogrammes de la figure 1 comparent les données (12 premières appels) à une distribution géométrique où le paramètre π est de 0,225. La concordance est assez grande. La statistique d'ordre d'échantillon semble indiquer que $\pi \in (0,2, 0,25)$. L'histogramme des données effectives d'enquête révèle que'il y a moins de sujets joints au premier appel qu'au deuxième. Il est possible qu'on ait fait plus de deuxièmes appels à un moment de la journée où le taux de succès était supérieur.

Supposons que $\pi = 0,225$. Selon la propriété « sans mémoire » d'une distribution géométrique, nous pouvons nous attendre à ce que 218 des 969 non-répondants aient effectivement répondu au téléphone à la 13^e tentative. Ainsi, les données sur les 13 premières tentatives d'entrée en communication seraient celles de la figure 2. Nous pouvons nettement voir que cette figure n'a pas le comportement d'une variable aléatoire suivant la distribution géométrique. La question est la suivante : si on avait appelé tous les sujets un nombre illimité de fois, les aurait-on tous joints? Si on répond oui à la question pour l'ensemble de données, on a le problème illustré à la figure 2.

réaction » par les variables indicatrices de deux des trois catégories, l'effet de la troisième catégorie étant alors absorbé dans le terme « ordonnée à l'origine ». Pour l'état de fumeur, les variables indicatrices « S_1^i » et « SQ_1^i » désignent respectivement l'usage actuel et antérieur du tabac. Pour l'état de réaction, les variables indicatrices « $b.USUL_1^i$ » et « $b.NO_1^i$ » désignent respectivement une réaction habituelle et une absence de réaction du sujet i à la fumée secondaire.

Soit X_i^j = le vecteur des variables explicatives pour le sujet i ,

$$X_i^j = (K-risk_i^j, S_i^j, SQ_i^j, b.USUL_i^j, b.NO_i^j, Age_i^j).$$

Par un modèle logistique multinomial non ordonné, nous considérons $p_j(x_i) = P(X_i^j = x_i)$, c'est-à-dire les probabilités que le sujet i réponde dans la catégorie $j \in \{0, 1, 2\}$, étant donné les variables explicatives observées pour ce sujet. Bien sûr, ce modèle utilise des équations linéaires η_{ij} décrivant en expression logarithmique les probabilités de réponse du sujet i dans la catégorie j par rapport à la catégorie de référence $j = 0$. Ainsi, pour $j = 1, 2$, nous voulons examiner :

$$(2) \quad \ln \frac{p_j(x_i)}{p_0(x_i)} = \eta_{ij} = \beta_{0j} + X_i^j \beta_j,$$

avec $\eta_{i0} = 0$. Les deux équations linéaires résultantes, η_{i1} et η_{i2} , ont chacune sept coefficients, soit une ordonnée à l'origine β_{0j} et les coefficients suivants :

$$\beta_j = (\beta_{K-risk_j^j}, \beta_{S_j^j}, \beta_{SQ_j^j}, \beta_{b.USUL_j^j}, \beta_{b.NO_j^j}, \beta_{Age_j^j}).$$

Le modèle de régression logistique MAR comporte 14 paramètres, le vecteur de ces 14 paramètres, représenté par $\beta = (\beta_{01}, \beta_{11}, \beta_{21}, \beta_{02}, \beta_{12}, \beta_{22})$, a la vraisemblance (ou, plutôt, la pseudo-vraisemblance puisque les poids sont incorporés au moyen de la variable W_i^j) :

$$(3) \quad L(\beta) \propto \prod_{j=0}^2 \prod_{i=1}^m \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{x_{ij}^j W_i^j}.$$

4.2 Régression logistique bayésienne

Nous reprenons la valeur de vraisemblance de l'équation (3) et les données recueillies auprès des répondants dans une analyse bayésienne. Les quatre variables explicatives de l'analyse fréquentiste précédente sont aussi reprises. Nous avons attribué des distributions *a priori* (voir la section 6) aux paramètres de régression logistique. Nous procédons par simulation MCMC pour la distribution postérieure des paramètres.

fonction des valeurs de cette variable, le mécanisme est dit « non-ignorable » (NI). Groves et Couper (1998) font observer que, si les probabilités de participation sont fonction de la variable dépendante visée, le biais de non-réponse peut être relativement important même avec un bon taux de réponse.

Soit R_i un indicateur de réponse, $R_i = I_{\text{répondant}}^{(i)}$ (sujet i) et $R = (R_1, \dots, R_n)^T$. Little et Rubin (1987) font voir qu'une méthode possible de prise en compte du mécanisme de non-réponse est l'inclusion dans le modèle de cette variable indicatrice de réponse. On peut qualifier le mécanisme de non-réponse d'ignorable si π et θ sont distincts et que :

$$f(R | Y^{\text{obs}}, Y^{\text{mis}}, \pi) = f(R | Y^{\text{obs}}, \pi) \quad (1)$$

où Y^{obs} et Y^{mis} représentent respectivement les données observées et les données manquantes de la variable dépendante d'intérêt.

Dans toute cette analyse, nous emploierons les termes « hypothèse MAR » et « hypothèse NI ». Précisons que, dans le premier cas, il s'agit de l'hypothèse du caractère non-informatif du mécanisme de non-réponse aux fins des inférences relatives à la variable dépendante indiquée à la section 2. En d'autres termes, les valeurs observées de cette variable constituent un sous-échantillon aléatoire de la population, peut-être à l'intérieur de strates, et il est inutile de tenir compte du mécanisme de non-réponse. Dans le second cas, il s'agit de l'hypothèse du caractère non-ignorable de ce même mécanisme et de l'impossibilité de traiter comme sous-échantillon aléatoire les données recueillies pour la variable dépendante. Plus précisément, les inférences au sujet de la population doivent tenir compte du mécanisme de non-réponse.

L'évaluation de l'hypothèse MAR se fait en trois étapes. On examine d'abord ce qu'on peut faire si on retient cette hypothèse. Comme la variable dépendante d'intérêt comporte trois catégories et que certaines des variables explicatives sont quantitatives, on recourt à une régression fréquentiste et la forme bayésienne d'un tel modèle.

On élabore ensuite un modèle NI. On modélise le mécanisme de non-réponse en se reportant au nombre de tentatives d'entrée en communication avec les divers sujets. Ici, on examine l'idée d'une fraction de « survivants » dans l'échantillon pour juger s'il est effectivement possible de joindre tous les sujets sélectionnés. Ensuite, on relie le mécanisme de non-réponse à la variable dépendante en intégrant le nombre d'appels au modèle de régression logistique.

Dans l'élaboration du modèle NI, nous employons une technique bayésienne de détermination des valeurs probables des données manquantes compte tenu des données d'observation et des paramètres du modèle. Nous appliquons à cette fin la technique d'augmentation de données où il y a imputation des données manquantes à chaque itération d'une simulation de Monte Carlo à chaîne de

Nous évaluons enfin l'hypothèse MAR. L'existence de coefficients non nuls pour le nombre d'appels dans le volet « régression logistique » du modèle NI implique que le nombre d'appels fait toute la différence, c'est-à-dire que les opinions de ceux qui n'ont pas répondu aux 12 premières tentatives d'entrée en communication sont susceptibles de différer des opinions de ceux qui ont répondu après quelques tentatives seulement. Dans ce cas, le mécanisme de non-réponse n'est pas indépendant des valeurs des données manquantes et une hypothèse MAR n'a pas sa place. Il s'agit ensuite d'examiner les probabilités de réponse en expression logarithmique pour les trois modèles. Les différences dégagées indiquent l'ordre de grandeur de l'erreur causée par une fausse hypothèse MAR. Ainsi, dans l'évaluation de ces hypothèses, on répond à deux questions : y a-t-il une différence? quelle est l'importance de cette différence?

4. MODÈLES MAR

4.1 Régression logistique

En nous reportant aux données recueillies auprès des $(m = 1\,429)$ sujets qui ont répondu à l'enquête, nous modélisons par une régression logistique pondérée les opinions de la population au sujet de l'usage du tabac en milieu de travail. Nous avons resserré l'échantillon des prédicteurs possibles (selon les questions de l'enquête et les renseignements généraux) par une série de tests de Wald. Nous avons ensuite comparé les modèles possibles par des tests de rapport des vraisemblances, CIA et CIB. Nous avons jugé que le modèle le mieux ajusté était celui qui comprenait des termes additifs pour les variables « Connaissance des risques », « Etat de fumeur », « Etat de réaction » et « Âge » (voir la section 2).

Comme chacun des modèles de notre analyse comporte un volet « régression logistique », il serait bon de mieux décrire ici la notation employée. La catégorie 0 « permission de fumer seulement dans des zones réservées » est notre catégorie de référence. On doit se rappeler que $Y_i \in \{0, 1, 2\}$. Pour le modèle MAR, nous nous reportons uniquement aux valeurs observées des opinions des sujets sur l'usage du tabac en milieu de travail, $X^{\text{obs}} = (X_1^{\text{obs}}, \dots, X_m^{\text{obs}})$. Soit $X_i^{(j)} = I^{(j)}(X_i)$ un indicateur du sujet i répondant dans la catégorie j et soit W_j le poids attribué à chaque sujet. Comme dans les analyses originales publiées de cet ensemble de données (Pederson et coll. (1996)), il y a pondération de ménage (voir Northrup (1993)) et de poststratification (voir l'annexe A) dans tous les modèles considérés ici.

Nous avons intégré à notre modèle les deux variables explicatives catégoriques « Etat de fumeur » et « Etat de

2. ENQUÊTE

Les données portent sur les réponses à 50 questions et sur 18 autres variables de caractérisation des sujets. Voici quelques indicateurs employés :

- « Connaissance des risques » est un résultat entier variant de 0 à 12 pour la connaissance des risques et des effets du tabagisme passif;
- « Etat de fumeur » indique si le sujet fume actuellement (S), a déjà fumé (SQ) ou n'a jamais fumé (NS);
- « Etat de réaction » indique si la fumée secondaire dérange le sujet : « dérange toujours » (b.A), « dérange d'habitude » (b.USUL) et « ne dérange pas » (b.NO);
- « Age » : (âge en années - 50) / 10.

Pederson, Bull, Ashley et Lefcoe (1989) ont élaboré un indicateur de connaissance des effets sur la santé du tabagisme passif à l'aide des réponses à six questions de l'enquête où on mesurait la connaissance qu'avait le sujet des effets de la fumée secondaire. Avec les questions de Pederson et coll., on a créé pour les données de phase III l'indicateur ici rebaptisé « Connaissance des risques ». Une note plus élevée pour la connaissance des risques indique que le sujet connaît mieux les dangers du tabagisme passif. Nous avons enfin modifié et remis à l'échelle la variable « Age » pour ainsi nous aligner sur le traitement de l'âge par Bull (1994) dans l'analyse des données des phases I et II.

3. APERÇU DE LA MÉTHODOLOGIE

Notre question fondamentale est la suivante : pouvons-nous ne pas tenir compte de la non-réponse par unité et traiter les données observées comme un sous-échantillon aléatoire de la population? Pour reprendre les termes de Little et Rubin (1987) et de Rubin (1976), s'il est possible de traiter les données d'observation de la variable dépendante d'intérêt comme un sous-échantillon aléatoire, les données manquantes sont alors entièrement aléatoires (« missing completely at random » ou MCAR). S'il est possible de traiter les données d'observation de la variable dépendante d'intérêt comme sous-échantillon aléatoire en conditionnant par les variables explicatives, les données manquantes sont « simplement aléatoires » (« missing at random » ou MAR). Si θ représente les paramètres des données et π , les paramètres du processus générant les données manquantes, Rubin (1976) dit de ces paramètres qu'ils sont distincts si on ne peut lier *a priori* π à θ par des restrictions d'espace paramétrique ni des distributions *a priori* de paramètres. Si le traitement MCAR ou MAR s'applique et que π et θ sont distincts, le mécanisme qui cause la non-réponse est jugé « ignorable » aux fins des inférences au sujet de la distribution de la variable d'intérêt. Si les données manquantes pour la variable dépendante sont

Dans la municipalité de Toronto, un règlement sur l'usage du tabac en milieu de travail est entré en vigueur le 1^{er} mars 1988. Depuis janvier de cette même année, on a effectué six enquêtes pour évaluer les attitudes de la population à l'égard du tabagisme, la sensibilisation au danger pour la santé de l'usage du tabac et l'incidence du règlement sur les résidents de la région métropolitaine de Toronto. Les données de notre analyse viennent de la troisième de ces enquêtes. Northrup (1993) livre des indications techniques sur cette dernière. Pour plus de clarté, précisons que nos données d'analyse sont celles de la troisième reprise de l'enquête et que les données des deux premières sont alors celles des phases I et II.

Northrup (1993) indique que les données d'intérêt, qui ont été mises à la disposition des intéressés par l'Institut for Social Research (ISR) de l'Université York, ont été recueillies auprès de 1 429 résidents de la région métropolitaine de Toronto en décembre 1992 et mars 1993. Aux fins de l'enquête, il y a eu sélection probabiliste en deux degrés des répondants. Au premier degré, on a fait de la « composition aléatoire » et, au deuxième degré, on s'est reporté au jour de naissance le plus récent pour sélectionner un adulte après entrée en communication avec le domicile admissible. On a alors pondéré les réponses selon le nombre d'adultes dans les ménages. Dans l'analyse qui suit, nous avons également appliqué une poststratification par groupe d'âge-sexe selon les données du recensement en vue d'une correction de sous-représentation de sous-populations. Au stade de la collecte de données, le nombre de lignes téléphoniques des ménages n'a pas été pris en considération. Le nombre de tentatives d'entrée en communication figure comme variable dans l'ensemble de données. Il n'y a pas de valeurs manquantes pour cette variable. Northrup (1993) explique que les 1 429 réponses ont été tirées d'un échantillon de 5 702 numéros de téléphone générés par la technique de composition aléatoire. Sur ces 5 702 ménages, 2 286 ont été jugés admissibles et 3 150, inadmissibles après vérification. On n'a pu déterminer l'admissibilité des 266 ménages restants. L'ISR a supposé que le taux d'admissibilité des ménages était le même pour ces 266 numéros de téléphone que pour le reste de l'échantillon de composition aléatoire. Ce taux implique un total estimatif de 2 398 ménages échantillonnés et un taux de réponse de 60 %. Ainsi, on estime à 969 le nombre de ménages sélectionnés qui n'ont pas répondu. Chacun a reçu 12 appels au minimum le jour, le soir et le week-end avant d'être classé comme « ménage non répondant ».

Aux fins de la présente analyse, la variable dépendante est l'option d'une personne sur la réglementation de l'usage du tabac en milieu de travail selon une des trois catégories suivantes : la catégorie 0 correspond à la permission de fumer seulement dans des zones réservées, la catégorie 1, à l'interdiction totale du tabac, et la catégorie 2, à une permission totale. Pour chaque sujet soumis à l'enquête, soit $Y_i \in \{0, 1, 2\}$ l'option du sujet i.

Effet de l'intensité des efforts en vue de joindre les répondants : enquête torontoise sur le tabagisme

LOUIS T. MARIANO et JOSEPH B. KADANE¹

RÉSUMÉ

Dans une enquête téléphonique, le nombre d'appels est utilisé comme indicateur de la difficulté à joindre le répondant. Ceci permet, dans un modèle de non-réponse, une division probabiliste des non-répondants en deux catégories : réfractaires (ceux qui refuseront toujours de répondre) et non-réfractaires (ceux qui ne sont pas disponibles pour répondre). Cela permet en outre d'estimer stochastiquement les opinions de ce dernier groupe de non-répondants et d'évaluer si la non-réponse est ignorable aux fins des inférences au sujet de la variable dépendante. Nous avons appliqué ces idées aux données d'une enquête dans la région métropolitaine de Toronto ayant porté sur les attitudes à l'égard de l'usage du tabac en milieu de travail. À l'aide d'un modèle bayésien, nous échantillonons la distribution postérieure des paramètres du modèle par les méthodes de Monte Carlo à chaîne de Markov. Les résultats révèlent que la non-réponse n'est pas ignorable et que ceux qui n'ont pas répondu étaient deux fois plus susceptibles d'accepter le libre usage du tabac en milieu de travail que ceux qui ont répondu.

MOTS CLÉS : Rappels, nombre de; analyse bayésienne; méthode de Monte Carlo à chaîne de Markov; non-réponse informative; non-réponse ignorable.

1. INTRODUCTION

Compte tenu des réalités de la non-réponse dans toute enquête, il est bon de juger comment on tiendra compte de cette dernière dans l'interprétation des données recueillies. Rubin (1976) énonce des conditions nécessaires et suffisantes pour qu'une telle analyse se confonde, des points de vue fréquentistes, de vraisemblance et bayésiens respectivement, avec une analyse reposant sur un modèle qui comporte un mécanisme de données manquantes. C'est en s'appuyant sur ce cadre que Little et Rubin (1987) ont enrichi une vaste documentation spécialisée d'une modélisation « informative et non-ignorable » de la non-réponse. En concernant l'interaction enquête-enquête, on peut affiner l'analyse de l'importance des données manquantes dans une enquête. Pour notre propos, nous citerons l'exemple d'une enquête sur les attitudes des Torontois à l'égard de l'usage du tabac en milieu de travail. On avait choisi des numéros de téléphone au hasard et, pour joindre les gens ainsi visés, on avait fait au moins 12 tentatives d'entrée en communication. Les données relatives aux répondants nous renseignent uniquement sur le nombre d'appels jusqu'à achèvement d'interview et ne précisent pas les moments où les tentatives infructueuses ont eu lieu. Même avec ces données moins riches sur la difficulté de joindre les répondants, nous constatons que le nombre d'appels infructueux est une indication importante au moment de considérer les résultats de l'enquête.

L'utilisation des données sur le nombre de tentatives d'entrée en communication avec l'enquête sélectionnée n'a rien d'unique. Porthoff, Manton et Woodbury (1993)

exposent une méthode de correction de biais d'enquête par indisponibilité qui prévoit une pondération en fonction du nombre de rappels. Notre analyse vise aussi le biais par indisponibilité, mais avec de grandes différences. Au lieu de supposer qu'il n'y a pas de refus, nous tenons compte de leur éventuelle existence dans une modélisation du mécanisme qui cause la non-réponse. Dans l'analyse qui suit, nous évaluons le rapport entre la non-réponse et la variable dépendante d'intérêt dans cette enquête avec les autres variables explicatives, et ce, après une pondération en fonction tant de la taille des ménages que de caractéristiques démographiques appropriées de la population. Nous nous trouvons donc à nous demander non seulement s'il y a erreur par indisponibilité, mais aussi si une stratification des répondants selon la taille des ménages et la structure âge-sexe actuelle peut écartier la nécessité de prendre l'erreur en compte par un mécanisme de description de la non-réponse. À noter que, dans ce cas, nous nous alignons sur les groupes de Pederson, Bull et Ashley (1996) dans les analyses originales publiées de l'ensemble de données. Des méthodes de correction de cellules plus complexes sont possibles (Little 1996; Eltinge et Yansaneh 1997, mentions bibliographiques de ces documents, etc.). Dans l'ordre de présentation de notre article, la section 2 renseigne plus en détail sur l'enquête, la section 3 expose la méthodologie employée, les sections 4 et 5 examinent respectivement les modèles « données manquantes aléatoires » et « données manquantes non-ignorables », la section 6 décrit les distributions *a priori* choisies pour l'analyse principale, la section 7 explique les résultats de cette analyse et la section 8 tire des conclusions.

¹ Louis T. Mariano est un candidat au doctorat, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; Joseph B. Kadane est Leonard J. Savage University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

BIBLIOGRAPHIE

BELINFANTE, A. (2000). Telephone Subscribcrshp in the United States. Industry Analysis Division, Common Carier Bureau, Federal Communications Commission, Washington, D.C. 20554.

BRICK, J.M., FLORES CERVANTES, I., WANG, K. et HANKINS, T. (1999). Evaluation of the use of data on interruptions in telephone service. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 376-381.

BRICK, J.M., WAKSBERG, J. et KEETER, S. (1996). Utilisation des données sur les interruptions du service téléphonique pour ajuster la couverture. *Techniques d'enquête*, 22, 187-199.

CURRENT POPULATION SURVEY (1978). Current Population Survey: Design and Methodology. Rapport technique 40. Department of Commerce, Bureau of the Census, Washington, D.C.

FRANKEL, M.R., EZZAT-RICE, T., WRIGHT, R.A. et SRINATH, K.P. (1998). Use of data in interruptions in telephone service for noncoverage adjustment. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 290-295.

FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. et SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 934-939.

KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

KISH, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics*, 8, 183-200.

LITTLE, R. et RUBIN, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons, 55-60.

LOHR, S. (1999). *Sampling: Design and Analysis*. New York: Duxbury Press, 255-287.

MASSEY, J., et BOTMAN, S. (1988). Weighting adjustments for random digit dialcd surveys. Dans *Telephone Survey Methodology*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls et J. Waksberg). New York: John Wiley and Sons, 143-160.

SMITH, T. (1990). Phone home? An analysis of household telephone ownership. *International Journal of Public Opinion Research*, 2, 369-390.

Codage des variables

0 - « Refus » nombre = 9
1 - 18 à 29 ans
2 - 30 à 44 ans
3 - 45 à 59 ans
4 - plus de 60 ans
Faible revenu
0 - Revenu du ménage supérieur à 20 000 \$ ou refus
1 - Revenu du ménage inférieur à 20 000 \$
Situation d'emploi
0 - Occupé à temps plein ou retraité
1 - Autre (refus, temps partiel, ménagère, étudiant, chômeur, autre)
Race
0 - Caucasicn, natif de l'Alaska, Hispanique ou Asiatique
1 - Amérindien, Afro-américain, Noir ou autre
Région appalachienne
0 - Ne vit pas dans un comté défavorisé du Kentucky, de l'Ohio ou de la Virginie occidentale
1 - Vit dans un comté défavorisé
Kentucky/ Virginie occidentale
0 - Ohio
1 - Kentucky ou Virginie occidentale

Résultats

Variable	B	E-T
Âge (refus)	-2,107	12,160
Âge (18 à 29)	2,006	0,357
Âge (30 à 44)	1,664	0,347
Âge (45 à 59)	1,064	0,364
Faible revenu	1,358	0,189
Situation d'emploi	0,397	0,187
Race	1,136	0,292
Région appalachienne	0,531	0,196
Kentucky/Virginie occid.	0,567	0,216
Constante	-5,712	0,401

Test de la qualité de l'ajustement de Hosmer Lemeshow

Chi carré	3,568
Degrés de liberté	8
Valeur p	0,894

Courbe ROC

Aire sous la courbe	0,782
---------------------	-------

compromis biais-variance.
Brick et coll. (1996) font remarquer que ces ajustements de la pondération pour tenir compte de la couverture téléphonique devraient représenter un meilleur moyen de

De nombreux éléments doivent être pris en considération pour déterminer quel scénario d'ajustement est le meilleur. Comme nous l'avons mentionné plus haut, les scénarios d'ajustement de la pondération par catégorie BWKE et BWT sont difficiles à appliquer si la population cible est très spécifique. Cependant, ces scénarios sont assez prudents, en ce sens qu'ils réduisent habituellement le biais sans augmenter la variance. Les scénarios fondés sur l'ajustement proportionnel itératif donnent ordinairement de meilleurs résultats que les scénarios d'ajustement de la pondération par catégorie, mais causent parfois une augmentation de la variance dans le cas d'ajustements importants de la pondération. Nous conseillons par conséquent de calculer les estimations suivant plusieurs scénarios, puis de déterminer lequel offre le meilleur

La méthode d'ajustement proportionnel itératif (*raking ratio*), TRAK, produit plusieurs valeurs estimatives très favorables du RQM. Ce scénario nous a permis de tenir compte de la différence de taux de pénétration du téléphone selon la région, mais non des différences selon d'autres caractéristiques démographiques. La variabilité a augmenté lorsque nous avons estimé les taux régionaux d'après les taux au niveau de l'état. Par conséquent, comme pour l'ajustement de la pondération par catégorie, le scénario donne de meilleurs résultats pour les échantillons de grandes populations. Bien que les valeurs moyennes et médiane du RQM estimé soient faibles pour ce scénario, certaines valeurs du rapport sont élevées. Les plus élevées sont celles observées pour l'Ohio où la proportion d'abonnés irréguliers dans l'échantillon est faible comparativement à la proportion estimative d'abonnés permanents. À lui seul, l'ajustement selon le score de propension, PROPM, réduit tout peu le biais pour être valable. Cependant, l'ajustement d'après le score de propension est avantageux parce qu'il permet de tenir compte des différences de probabilité d'être abonné au service téléphonique sans devoir utiliser des données extérieures. Conjugué à l'ajustement proportionnel itératif, le scénario fondé sur le score de

Les estimations types. Le tableau 5 montre que les scénarios BWKE et BWKT produisent la plupart du temps une meilleure estimation. Souignons aussi que, même quand ces scénarios produisent une estimation qui n'est pas meilleure, l'augmentation de la variance reste assez faible. La méthode d'ajustement de la pondération par catégorie donne de bons résultats pour les échantillons de grande population, tels que les états ou les pays, puisque les données auxiliaires nécessaires pour calculer les ajustements peuvent être obtenues facilement. La méthode est plus difficile à appliquer pour des échantillons de popula-

ANNEKE

Régression logistique de l'état d'abonné irrégulier

Suit la description du modèle que nous avons utilisé pour prédire l'état d'abonné irrégulier. La plupart des variables du modèle ont trait au statut socioéconomique. Les coefficients indiquent que les personnes jeunes, les personnes à faible revenu, celles qui ne sont pas occupées à temps plein, les Américains et les Afro-américains, et les habitants des comtés défavorisés ont une plus forte propension que les autres à être abonnés irrégulièrement au téléphone. Le niveau de signification élevé du test de Hosmer et Lemeshow indique que l'ajustement du modèle est excellent. L'aire importante sous la courbe ROC nous indique que le pouvoir discriminatif du modèle est bon.

REMERCIEMENTS

L'estimation, qui figure au dénominateur, diminue. Les résultats de cette étude et d'autres indiquent que les ajustements pourraient être utiles pour bon nombre d'estimations calculées d'après les données d'enquêtes téléphoniques et devraient être pris en considération sérieux-ment. Les avantages de l'ajustement semblent surpasser ses inconvénients dans le cas des scénarios d'ajustement de la pondération par catégorie, d'ajustement proportionnel itératif et du score de propension augmenté. Compte tenu de la petite taille de l'échantillon et de la population spéciale visée par l'Appalachian Poll, il est conseillé de ne pas généraliser ces résultats tant que les méthodes n'auront pas été évaluées plus en détail. Elles doivent encore être testées en se fondant sur une enquête dépourvue de biais de couverture, c'est-à-dire une enquête dont la base de sondage comprend les ménages non abonnés au téléphone et qui fournit des renseignements sur la situation d'abonne-ments au téléphone, afin d'évaluer la validité des hypo-thèses. Les données de la National Survey of America's Families ou de la National Health Interview Survey pourraient convenir pour évaluer les méthodes d'ajustement et les hypothèses.

réduire l'erreur quadratique moyenne si l'échantillon d'enquête est de grande taille. Le rapport des biais augmente à mesure que la taille de l'échantillon augmente, puisque le biais ne varie pas, mais que l'erreur type de

biais. En l'absence d'estimations de référence non biaisées, il est impossible de valider cette hypothèse. Les rapports quadratiques moyens présentés ici sont vraisemblablement biaisés par défaut, puisque le biais qui entache l'estimation ajustée n'est pas inclus. Les RQM estimatifs sont néanmoins utiles pour comparer les méthodes et donner une bonne idée de l'efficacité des ajustements de la pondération.

Comme prévu, nous constatons que la méthode du nombre de jour DAY donne lieu à une trop forte variabilité pour être utile. La méthode des groupes de jour (DAYG) semble donner de meilleurs résultats, mais la plupart des rapports quadratiques moyens s'approchent de 100, ce qui signifie que l'amélioration par rapport à l'estimation type

n'est pas importante. L'avantage de ce scénario tient à sa simplicité. L'ajustement de la pondération est facile à appliquer et ne nécessite pas de données auxiliaires. Les scénarios d'ajustement de la pondération par catégorie ont l'avantage d'accorder plus de poids aux répondants compris dans les cellules où la probabilité d'avoir un téléphone est faible. Pour ces scénarios, la réduction du biais est plus forte pour les variables corrélées aux variables de classification. Par exemple, la propriété du logement et la propriété d'un ordinateur sont des variables positivement corrélées et le scénario BWKT, où les répondants sont classés selon la propriété du logement, produit des estimations de la proportion de ménages possédant un ordinateur systématiquement plus faibles que

Tableau 5
Rapport quadratique moyen pour certaines caractéristiques

Caractéristique	Rapport quadratique moyen des FIV									
	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP	DAY	DAYG	BWKE
Propriétaire	96,1	97,2	97,3	88,1	87,5	99,8	84,5	98,6	98,2	88,4
Non appal, Ohio	42,3	97,4	98,7	68,9	57,9	99,1	52,5	71,7	96,5	89,6
Appal, Ohio	69,9	97,6	98,7	83,1	77,7	99,4	83,5	71,7	98,2	89,6
Non appal, Kentucky	89,3	96,1	102,3	77,0	110,6	100,3	112,4	116,0	100,7	104,1
Appal, Kentucky	16,6	71,2	89,1	62,3	75,8	99,2	75,5	28,6	81,1	94,4
Non appal, Virginie occid.	20,2	98,4	103,2	100,3	109,3	100,5	110,6	43,5	106,0	101,1
Appal, Virginie occid.	99,9	99,0	88,2	61,4	51,9	99,5	48,5	98,8	100,5	112,1
Pas d'ass.-maladie	126,8	101,3	102,1	106,5	125,1	99,9	123,5	92,3	98,8	95,6
Appal, Ohio	206,4	99,9	100,9	102,8	103,0	99,8	102,7	39,0	87,9	90,7
Non appal, Kentucky	82,7	106,7	104,4	103,7	102,1	97,9	84,3	53,5	109,9	104,9
Appal, Kentucky	100,2	97,1	90,6	84,0	77,7	97,9	71,3	136,6	99,2	94,0
Non appal, Virginie occid.	149,6	99,2	74,1	83,5	52,8	95,7	46,7	107,0	96,5	75,1
Appal, Virginie occid.	86,5	90,5	80,5	57,9	45,2	99,7	45,6	105,2	100,8	104,3
Non appal, Ohio	31,9	92,9	97,4	86,1	48,6	99,4	46,4	68,5	98,2	96,8
Appal, Ohio	139,1	94,1	78,2	69,5	69,5	96,7	64,3	320,7	96,8	91,9
Non appal, Kentucky	22,3	55,8	57,6	58,7	31,0	97,2	31,9	30,5	68,4	68,5
Appal, Kentucky	117,3	102,6	82,4	71,0	59,6	97,7	57,2	105,7	101,9	94,7
Non appal, Virginie occid.	181,6	97,0	84,1	88,5	50,4	94,1	39,9	92,2	98,8	89,6
Appal, Virginie occid.	98,1	98,5	96,4	88,2	88,1	99,8	86,1	99,5	102,0	102,1
Non appal, Ohio	127,2	101,2	103,1	101,2	96,2	99,7	92,5	116,0	99,6	101,2
Appal, Ohio	67,7	94,9	94,4	92,7	93,6	99,5	92,8	27,1	93,7	91,5
Non appal, Kentucky	147,1	89,0	91,7	85,1	76,1	98,5	73,5	59,6	95,8	85,1
Appal, Kentucky	66,8	96,9	86,6	85,8	75,1	98,5	73,5	59,6	95,8	85,1
Non appal, Virginie occid.	82,7	95,6	104,4	103,0	111,2	99,6	108,2	41,8	88,1	101,6
Appal, Virginie occid.	137,6	97,5	94,3	92,2	85,2	99,3	83,8	125,2	99,1	97,1
Moyenne	107,5	99,0	99,1	97,1	89,8	99,8	86,3	94,8	98,9	98,7
Médian	109	55,8	0,9	57,9	4,1	94,1	5,7	7,0	68,4	43,1
Minimum	607,7	108,5	104,8	104,8	109,1	133,1	100,5	695,2	140,8	144,5
Maximum	47,1	60,8	61,8	61,8	58,8	65,7	87,3	63,7	62,7	56,9
Pourcentage en dessous de 100	58,8	58,8	54,4	54,4	58,8	58,8	58,8	58,8	58,8	58,8

Nota : En plus des quatre proportions énumérées dans le tableau, le sommaire des 17 variables inclut les variables suivantes : inquiet au sujet du revenu, meilleure situation économique durant les années 1990, insatisfait de son propre avoir, marié, a des enfants, chômeur, diplôme collégial, en bonne ou en excellente santé, maladie grave dans le ménage, pas de médecin de famille, satisfait de son propre logement, eau potable très sûre et accès à Internet à domicile.

où $\text{var}_i(p)$ est la variance estimative de l'estimation ajustée, obtenue par répétition. Un RQM de 100 indique que la variance de l'estimation ajustée est exactement égale à l'erreur quadratique moyenne de l'estimation type. Un RQM supérieur à 100 signifie que la variance de l'estimation ajustée est plus grande que l'EQM de l'estimation type et que le compromis biais-variance pour le scénario étudié n'est pas favorable. Un RQM inférieur à 100 signifie que l'estimation ajustée est meilleure que l'estimation type en ce qui concerne l'erreur globale.

Le tableau 5 donne les valeurs estimatives du RQM pour certaines variables d'enquête de l'Appalachian Poll, ainsi qu'un sommaire de ces valeurs pour 17 variables pour chaque scénario d'ajustement. Les estimations du RQM varient selon la région et selon le scénario. Les valeurs du RQM calculées au moyen de deux estimations différentes de la variance diffèrent aussi, mais les valeurs sommas des sont comparables pour les deux variantes. Le scénario DAY est celui qui donne les valeurs du RQM les plus élevées, ce qui laisse entendre que cette méthode d'ajustement de la pondération n'est pas judicieuse, car elle augmente trop la variance. Les scénarios TRAK et AUGP sont ceux qui produisent les valeurs moyennes et médianes du RQM les plus faibles, mais ils donnent des estimations défavorables pour quelques caractéristiques, comme l'indiquent les valeurs maximales élevées du RQM. Les scénarios d'ajustement de la pondération par catégorie BWKE et BWT donnent de bons résultats et la valeur estimative maximale du rapport quadratique moyen ainsi obtenue est assez faible. Pour le scénario PROF, toutes les valeurs du RQM s'approchent de 100, ce qui donne à penser que l'erreur globale des estimations calculées selon ce scénario est comparable à l'erreur des estimations.

5. CONCLUSIONS

Bien que l'usage du téléphone soit très répandu, les enquêtes téléphoniques seront toujours entachées d'un certain biais, puisque les ménages non abonnés sont exclus de la base de sondage et que les caractéristiques de la population de non-abonnés diffèrent de celles de la population d'abonnés. Le biais de couverture peut être réduit par post-stratification sur des variables telles que le revenu et le niveau de scolarité et n'est pas nécessairement problématique dans tous les cas. Cependant, lorsque les enquêtes visent des régions pauvres ou rurales où le taux de pénétration est faible, le biais de couverture pose un problème important.

Nous proposons quelques nouvelles méthodes pour réduire le biais de couverture par ajustement des coefficients de pondération attribués aux répondants faisant partie de la population d'abonnés irréguliers. Nous comparons les estimations résultantes à celles produites par d'autres méthodes existantes. Pour comparer ces méthodes, nous supposons que les estimations ajustées sont dépourvues de

se nourrir et un moins grand nombre possèdent un ordina-leur à la maison selon les estimations ajustées que selon les estimations obtenues par pondération type. Pour l'assurancemalade, la direction du biais est en grande partie uniforme pour les diverses régions. L'estimation type est entachée d'un biais par excès pour la région appalachienne de l'Ohio et la région non appalachienne du Kentucky et est, en général, entachée d'un biais par défaut pour les autres régions.

La grandeur absolue de la réduction du biais n'est pas, en soi, entièrement significative, car elle ne tient pas compte de l'importance de l'erreur d'échantillonnage associée à l'estimation. Par conséquent, à l'exemple de Brick et coll. (1996), nous calculons aussi le rapport des biais. Ce rapport pour le scénario i , r_i , est donné par

$$r_i = \frac{b_i}{\text{se}(\hat{p}_i)} \tag{5}$$

où $\text{se}(\hat{p}_i)$ représente l'erreur-type de l'estimation type. Le tableau 4 donne aussi le rapport des biais pour les estimations présentées. DAY, TRAK et AUGP donnent les rapports des biais les plus importants; pour ces scénarios d'ajustement, le biais n'est pas négligeable si nous considérons l'erreur-type. Pour DAYG et PROF, le rapport des biais est faible, donc la réduction du biais est faible comparativement à l'erreur de l'estimation.

4.3 Erreur quadratique moyenne

Puisque nous considérons que les estimations types sont biaisées, nous devrions évaluer l'erreur d'après l'erreur quadratique moyenne plutôt que la variance. L'EQM de l'estimation type correspond approximativement à

$$\text{eqm}_i = \text{var}(\hat{p}_i) + b_i^2 \tag{6}$$

Pour chaque scénario d'ajustement. Rappelons que nous supposons que les estimations ajustées sont dépourvues de biais, si bien que leur erreur quadratique moyenne doit être égale à leur variance. Nous pouvons calculer selon deux méthodes la valeur approximative de la variance des estimations ajustées. Pour obtenir la première approximation, nous multiplions le rapport des FTV du tableau 3 par la variance de l'estimation type. Nous pouvons aussi estimer la variance des estimations ajustées par des méthodes de répétition.

Pour comparer l'erreur de l'estimation ajustée à l'erreur moyen (RQM). Si nous utilisons la variance FTV, le RQM estimé est donné par

$$\text{rqm}_{\text{FTV}_i}(\hat{p}) = \frac{\text{eqm}_i(\hat{p})}{100 \times \text{rapport des FTV}_i \times \text{var}(\hat{p}_i)} \tag{7a}$$

Pour la méthode par répétition de la variance, le RQM estimé est donné par

$$\text{rqm}_{\text{VAR}_i}(\hat{p}) = \frac{\text{eqm}_i(\hat{p})}{100 \times \text{var}_i(\hat{p})} \tag{7b}$$

Tableau 3
Rapport des facteurs d'inflation de la variance liés à l'ajustement des coefficients de pondération

Région	Rapport du FTV du scénario d'ajustement au FTV de la pondération type									
DAY	0,999	0,997	1,004	1,023	1,063	0,999	1,061	Non appalachienne, Ohio		
DAYG	1,016	1,018	1,019	1,054	1,030	0,999	1,336	Appalachienne Ohio		
DAYG	1,040	1,040	1,049	1,054	1,030	0,999	1,336	Non appalachienne, Kentucky		
DAYG	1,069	1,069	1,045	1,042	1,129	1,003	1,145	Appalachienne, Kentucky		
DAYG	2,433	2,433	2,433	2,433	2,433	2,433	2,433	Non appalachienne, Virginie occident.		
DAYG	2,935	2,935	2,935	2,935	2,935	2,935	2,935	Appalachienne, Virginie occident.		
DAYG	3,055	3,055	3,055	3,055	3,055	3,055	3,055	Moyenne des scénarios		
DAYG	1,039	1,029	1,049	1,049	1,115	1,001	1,190			
DAYG	1,039	1,029	1,049	1,049	1,115	1,001	1,190			

Tableau 4
Réduction estimée du biais et rapport des biais pour certaines caractéristiques

Propriétaire	Caractéristique	Estimation	Erreur- type	Estimation type	Réduction estimée du biais	Rapport des biais
DAY	DAYG BWKE	PROP	AUGP	DAY	DAYG BWKE	BWKT TRAK PROP AUGP

Propriétaire	Non appal., Ohio	72,2	3,1	0,6	0,5	0,5	1,2	1,4	0,1	1,6	0,2	0,2	0,2	0,2	0,2	0,4	0,5	0,0	0,5
	Appal., Ohio	75,4	2,8	4,4	0,6	0,6	2,1	3,2	0,3	3,5	1,6	0,2	0,2	0,2	0,2	0,3	0,6	0,1	1,2
	Non appal., Kentucky	68,6	3,1	7,2	0,8	0,9	1,8	1,5	0,2	1,5	2,3	0,3	0,3	0,3	0,6	0,5	0,1	0,5	0,5
	Appal., Kentucky	80,0	2,2	2,9	0,3	0,3	1,3	0,3	0,0	0,3	1,3	0,3	0,3	0,1	0,6	0,1	0,0	0,1	0,1
	Non appal., Virginie occid.	80,0	2,3	14,2	1,6	0,9	1,9	1,4	0,2	1,4	6,1	0,7	0,4	0,8	0,6	0,1	0,0	0,1	0,6
	Appal., Virginie occid.	81,9	2,2	8,2	0,7	-0,4	0,5	-0,3	0,0	-0,2	3,7	0,3	-0,2	0,2	-0,1	0,0	-0,1	-0,1	-0,1
Pas d'ass.-malade	Non appal., Ohio	7,3	1,7	0,0	-0,1	-0,6	-1,4	-1,7	-0,1	-1,8	0,0	-0,12	-0,4	-0,8	-1,0	-0,1	-1,1	-0,1	-1,1
	Appal., Ohio	12,6	2,1	0,9	0,1	0,3	0,3	0,5	0,1	0,6	0,4	0,1	0,1	0,2	0,3	0,0	0,3	0,3	0,3
	Non appal., Kentucky	8,8	1,8	1,8	0,4	0,2	0,3	0,0	0,1	0,1	1,0	0,2	0,1	0,2	0,1	0,0	0,0	0,0	0,0
	Appal., Kentucky	22,2	2,4	3,4	0,1	-0,1	-0,2	-0,4	-1,5	-1,4	1,4	0,0	0,0	0,1	-0,1	-0,3	-0,2	-0,7	-0,6
	Non appal., Virginie occid.	14,2	2,1	-4,8	-0,5	-0,7	-1,0	-1,2	-0,3	-1,4	-2,3	-0,2	-0,3	-0,5	-0,6	-0,1	-0,7	-0,7	-0,7
Pas suffisamment d'argent	Appal., Virginie occid.	24,6	2,5	2,5	-0,8	-1,7	-1,3	-2,7	-0,6	-3,0	1,0	-0,3	-0,7	-0,5	-1,1	-0,2	-1,2	-1,2	-1,2
pour se nourrir	Non appal., Ohio	10,8	1,9	-0,7	-0,6	-0,9	-1,6	-2,2	-0,1	-2,1	-0,4	-0,3	-0,5	-0,9	-1,2	0,0	-1,2	-1,2	-1,2
	Appal., Ohio	16,2	2,5	-4,7	-0,8	-1,3	-3,3	-3,3	-0,2	-3,4	-1,9	-0,3	-0,3	-0,5	-1,3	-0,1	-1,4	-0,8	-1,4
	Non appal., Kentucky	11,4	2,4	-3,3	-0,8	-1,3	-1,7	-1,6	-0,4	-1,8	-1,4	-0,3	-0,5	-0,7	-0,7	-0,1	-1,4	-0,8	-1,4
	Appal., Kentucky	20,2	2,4	-7,4	-2,3	-2,1	-3,8	-3,8	-0,4	-3,8	-3,1	-1,0	-0,9	-0,9	-1,6	-0,2	-1,6	-0,9	-1,6
	Non appal., Virginie occid.	14,0	2,1	4,3	-0,1	-1,0	-1,4	-1,7	-0,3	-1,8	2,1	0,0	-0,5	-0,7	-0,8	-0,2	-0,9	-0,9	-0,9
	Appal., Virginie occid.	16,4	2,0	1,5	-0,7	-1,0	-0,9	-2,2	-0,5	-2,6	0,8	-0,3	-0,5	-0,4	-1,1	-0,3	-1,3	-1,3	-1,3
Ordinateur à la maison	Non appal., Ohio	60,1	3,0	0,4	0,3	0,6	1,2	1,3	0,1	1,4	0,1	0,1	0,1	0,2	0,4	0,5	0,0	0,5	0,5
	Appal., Ohio	40,0	3,0	1,2	0,3	0,8	1,8	1,8	0,1	2,0	0,4	0,1	0,1	0,3	0,6	0,0	0,7	0,7	0,7
	Non appal., Kentucky	44,5	3,0	6,7	0,9	0,8	0,9	0,9	0,2	1,0	2,3	0,3	0,3	0,4	0,3	0,1	0,3	0,3	0,3
	Appal., Kentucky	29,7	2,3	1,9	1,0	0,9	1,1	2,3	0,0	1,9	0,8	0,4	0,4	0,5	1,0	0,0	0,8	0,8	0,8
	Non appal., Virginie occid.	46,2	2,6	7,6	0,6	1,1	1,2	1,5	0,3	1,6	2,9	0,2	0,4	0,4	0,6	0,1	0,1	0,6	0,6
	Appal., Virginie occid.	36,1	2,7	4,3	1,0	0,3	0,4	0,2	0,3	0,5	1,6	0,4	0,4	0,1	0,1	0,1	0,1	0,2	0,2
Sommaire des 17 variables																			
Valeur moy. absolue		0,032	0,005	0,006	0,009	0,013	0,002	0,014	0,001	0,014	0,995	0,240	0,235	0,260	0,412	0,605	0,055	0,075	0,885
Valeur méd. absolue		0,022	0,005	0,006	0,011	0,014	0,001	0,014	0,001	0,014	0,995	0,240	0,245	0,420	0,605	0,055	0,075	0,885	0,885

Nota : En plus des quatre proportions énumérées dans le tableau, le sommaire des 17 variables inclut les variables suivantes : inquiet au sujet du revenu, meilleure situation économique durant les années 1990, insatisfait de son propre avoir, mère, a des enfants, chômeur, diplômé(e) collégial, en bonne ou en excellente santé, maladie grave dans le ménage, pas de médecin de famille, satisfait de son propre logement, eau potable très sûre et accès à Internet à domicile.

entre l'estimation type et l'estimation ajustée. Nous obtenons sept estimations distinctes de la réduction du biais, une pour chaque scénario. La réduction estimée du biais est donnée par

biais est donnée par

$$(4) \quad \hat{d}^i d^s = q^i,$$

où b_i représente la réduction estimée du biais grâce au scénario i , \hat{p}_s représente l'estimation type et \hat{p}_i l'estimation produite par le scénario d'ajustement i . Le tableau 4 donne,

pour chaque scénario, les réductions estimées du biais pour quatre caractéristiques pour chacune des six strates. Pour les caractéristiques « propriétaire du logement », « pas assez d'argent pour se nourrir » et « ordinaire », la direction du biais est assez uniforme pour les divers scénarios et les diverses régions. Fait rassurant, cette direction est celle prévue pour ces caractéristiques, puisqu'un moins grand nombre de personnes sont propriétaires de leur logement, un plus grand nombre n'ont pas suffisamment d'argent pour

plus élevé est attribué aux ménages échantillonnés qui sont plus semblables que les autres aux non-répondants. Comme il n'existe généralement aucune donnée sur la population sans téléphone exclue des enquêtes téléphoniques, nous adoptons une méthode modifiée d'utilisation du score de propension. Nous ajustons uniquement les coefficients de pondération pour tenir compte des abonnés irréguliers puisque ce sont eux qui représenteront la part manquante de l'échantillon; les coefficients de pondération appliqués aux abonnés réguliers ne sont pas ajustés. L'ajustement du coefficient de pondération appliqué aux abonnés irréguliers est $1/(1-p)$, où p représente la propension estimative à l'abonnement irrégulier décrite par le modèle présentée à la section 2.1. Les ménages dont la propension estimée à être abonnés irrégulièrement au téléphone est forte pourraient être plus représentatifs que les autres de la population de non-abonnés au téléphone et se voient donc appliquer un ajustement de pondération plus important. L'ajustement est appliqué aux coefficients de pondération de base et le scénario est appelé propension (PROP).

L'abonnement irrégulier au téléphone n'est pas tellement courant et la plupart des scores estimatifs de propension sont assez faibles. Dans le scénario PROP, l'ajustement moyen du coefficient de pondération pour un ménage abonné irrégulièrement au téléphone est égal à 1,167. Cet ajustement n'est pas suffisamment important pour que les abonnés irréguliers soient représentatifs d'eux-mêmes et de la population entière de non-abonnés. Autrement dit, lorsque nous rééchantillons les coefficients de pondération de sorte que leurs sommes correspondent aux chiffres de population, la somme des coefficients finals pour les abonnés irréguliers est inférieure à la taille de la population de ces abonnés. Pour tenir compte de ce sous-dénombrement, nous appliquons l'ajustement de la pondération fondé sur le score de propension, puis nous utilisons la population d'abonnés irréguliers comme variable de contrôle pour l'ajustement proportionnel itératif, selon l'âge, le niveau de scolarité et le sexe. Les chiffres estimatifs de population pour les abonnés irréguliers sont présentés à la section 3.3. Nous donnons à ce scénario de pondération le nom de propension augmentée ou AUGP.

4. RÉSULTATS

L'analyse et la comparaison des scénarios d'ajustement présentées ici font pendant à l'analyse réalisée par Brick et ses collaborateurs (1996). Nous examinons pour commencer la variation de la variance due à l'ajustement des coefficients de pondération en vue de réduire le biais de couverture et nous présentons une statistique permettant de mesurer la variabilité relative. Puis, nous évaluons les scénarios en comparant la variance des estimations ajustées à l'erreur quadratique moyenne de l'estimation type.

4.1 Changement de variabilité

Le but des scénarios d'ajustement est de réduire le biais de couverture tout en contrôlant la variance. L'ajustement des coefficients de pondération pour réduire le biais augmente la variabilité de ces coefficients, donc augmente la variance des estimations. Kish (1992) donne une formule pour évaluer l'augmentation de la variance due à l'inégalité des coefficients de pondération. Brick et coll. (1996) qualifient cette expression de facteur d'inflation de la variance (FIV). Le FIV peut s'écrire sous la forme

$$\text{FIV} = 1 + [\text{CV}(\text{coeff. de pondération})]^2, \quad (3)$$

où CV (coeff. de pondération) est le coefficient de variation des coefficients de pondération. Nous calculons le rapport des FIV pour comparer le FIV d'un nouveau scénario de pondération à celui du scénario type. Le tableau 3 donne les rapports des FIV pour les six strates de l'échantillon de l'Appalachian Poll pour chaque scénario décrit à la section 3. Par exemple, un rapport des FIV de 1,12 indique que la variance augmente de 12 % par rapport à sa valeur lorsque l'on utilise le scénario de pondération type. La valeur du rapport des FIV est raisonnable pour tous les scénarios, sauf DAY pour lequel l'augmentation moyenne de la variance est de 300 %. Pour PROP, les valeurs du rapport des FIV sont toutes très proches de l'unité, ce qui donne à penser que les ajustements de la pondération selon ce scénario n'augmentera pas la variance des estimations.

4.2 Réduction du biais de couverture

Nous avons estimé 17 proportions de population en nous servant des variables d'enquête de l'Appalachian Poll suivant la méthode type de pondération et chacun des sept scénarios étudiés d'ajustement (voir le tableau 4 pour la liste des 17 variables). Nous nous sommes servis du logiciel WesVar pour calculer les erreurs-types de ces estimations par répétition. Nous aimerions évaluer l'efficacité de réduction du biais de couverture de chaque scénario en regard de ces 17 caractéristiques. Des estimations de pourcentage de tout biais de couverture téléphonique provenant d'une source indépendante représenteraient les données de référence idéales. Malheureusement, de telles données n'existent pas et certaines hypothèses doivent être formulées concernant le modèle afin de procéder à l'évaluation. Nous supposons que les méthodes d'ajustement des coefficients de pondération réduisent le biais de couverture. Donc, nous considérons la différence entre l'estimation type et l'estimation ajustée comme une estimation non biaisée de la diminution du biais de couverture due à l'ajustement. L'hypothèse favorable des estimations ajustées, puisqu'elles sont considérées comme étant dépourvues de biais.

Partant de notre hypothèse, nous comparons l'estimation produite par chaque scénario à l'estimation type. Nous estimons la réduction du biais de couverture par différences

service téléphonique. Par conséquent, les ajustements de la pondération par cellule calculées pour l'Appalachian Poll se fondent sur les données de la CPS agréées pour les trois états.

3.3 Ajustement proportionnel itératif (taking ratio)

Lohr (1999) explique l'utilisation d'estimations obtenues par ajustement proportionnel itératif pour faire la correction pour la non-réponse aux enquêtes. Nous proposons d'utiliser une méthode similaire pour tenir compte du biais de couverture. Nous estimons la proportion de la population abonnée en permanence au téléphone, puis nous procédons à un ajustement proportionnel itératif pour nous assurer que les abonnés irréguliers compris dans l'échantillon soient représentatifs de la part de la population qui n'est pas raccordée continuellement au service téléphonique.

Nous estimons la proportion de ménages non abonnés en permanence comme suit :

$$(2) \quad 1 - \left(\frac{\bar{t}_1 + \bar{t}_2}{\bar{t}_1 + \bar{t}_2 + \bar{t}_4} \right) \left(\frac{\bar{t}_1 + \bar{t}_2}{\bar{t}_1 + \bar{t}_2} \right),$$

où $\bar{t}_i, i = 1, 2, 4$, est déterminé d'après les données de la FCC. La première fraction donne une estimation de la proportion de ménages qui reçoivent le service téléphonique au moment de l'enquête et la deuxième, de la proportion de ménages abonnés régulièrement parmi les ménages qui reçoivent le service. De nouveau, nous supposons que $t_3 = 0$. La FCC donne les taux de pénétration du téléphone selon l'état, mais non selon la région. Les données du Recensement de 1990, quant à elles, donnent les taux de pénétration selon le comté, mais ces taux ont varié de 1990 à 1999. Par conséquent, pour estimer

Tableau 2
Calcul des totaux utilisés pour l'ajustement proportionnel itératif selon l'état d'abonné irrégulier

Données de l'Appalachian Poll		Kentucky		Ohio		Virginie occidentale	
		Ap.	Non ap.	Ap.	Non ap.	Ap.	Non ap.
Taille de l'échantillon	412	407	413	405	411	415	41
Nbre d'abonnés irrég. dans l'échant.	38	19	18	13	36	16	3
% de l'échant. sans service permanent	9,2	4,7	4,4	3,2	8,8	3,9	7,3
Données du recensement et de la FCC							
% sans téléphone, selon l'état, 1990	10,2	10,2	10,2	4,7	10,3	10,3	10,3
% sans téléphone, selon la région, 1990	19,1	8,2	11,7	4,5	14,3	8,4	7,3
% sans téléphone, selon l'état, 1999	6,7	6,7	5,2	5,2	7,3	68,2	7,3
% de la pop. de l'état dans la région	18,6	81,4	2,6	97,4	31,8	68,2	7,3
Estimations							
Ratio de non-couvert. Non ap./Ap.	0,429	0,429	0,385	0,385	0,587	0,587	0,587
% estim. sans service permanent	20,6	9,8	16,7	8,1	18,0	9,6	9,6
Nbre souhaité irrég. dans l'éch.	85	40	69	33	74	40	40

Dans un scénario que nous appelons ajustement itératif proportionnel selon l'état d'abonné irrégulier ou TRAK, la catégorie d'abonnés irréguliers est incluse à titre de variables de contrôle pour l'ajustement proportionnel itératif, en plus de l'âge, du sexe et du niveau de scolarité. Les totaux utilisés comme contraintes sur les marges pour l'ajustement proportionnel itératif selon l'état d'abonné irrégulier figurent au tableau 2.

3.4 Nouvelle pondération fondée sur les scores de propension

Un score estimatif de propension est parfois utilisé pour corriger la pondération afin de tenir compte de la non-réponse aux enquêtes pour lesquelles certaines caractéristiques des non-répondants sont connues. Par exemple, dans le cas d'une interview sur place, l'intervieweur connaît l'adresse du non-répondant et possède aussi des renseignements sur la race, le sexe et l'âge. Le cas échéant, on élabore un modèle de régression logistique qui décrit la propension à répondre et on attribue au répondant un coefficient de pondération, $1/p$, où p représente la propension estimative à répondre (Little et Rubin 1987). Selon cette méthode, un coefficient de pondération plus

habituellement déconseillé d'utiliser des coefficients de pondération dont la valeur est supérieure à trois. En fait, très souvent, dans le cas des grandes enquêtes, si les coefficients de pondération sont supérieurs à 2, le U.S. Census Bureau rassemble les répondants en groupes plus grands et calcule un coefficient de pondération de groupe pour obtenir des facteurs plus faibles d'ajustement de la pondération; consulter, par exemple, CPS (1978).

Cette méthode simple devient plus pratique si les répondants sont regroupés selon la durée de l'interruption du service. Dans un scénario que nous appelons groupes de jours (DAYG), les abonnés irréguliers sont regroupés en quartiles sur l'ensemble de l'échantillon, en fonction de la durée de l'interruption du service téléphonique. Ces quartiles correspondent à des interruptions d'une semaine, de plus d'une semaine mais de moins de trois semaines, de trois semaines à deux mois et de plus de deux mois. Pour chaque groupe, l'ajustement du facteur de pondération est donné par $365/(365 - \text{nombre moyen de jours sans service})$ et est de nouveau appliqué après le calcul du coefficient de pondération de base, mais avant l'ajustement proportionnel itératif. Cette méthode de groupement permet de réduire la variance due à des débordements de très longue durée.

3.2 Scénario d'ajustement de la pondération par catégorie

Brick et coll. (1996) appliquent aussi un ajustement de la probabilité de réponse afin de réduire le biais de couverture. Leur méthode consiste à répartir la population cible entre les quatre composantes décrites à la section 2 : t_1 représente le nombre de personnes qui vivent dans les ménages abonnés en permanence au téléphone, t_2 , le nombre de personnes qui vivent dans les ménages abonnés irrégulièrement mais qui reçoivent le service au moment de l'enquête, t_3 , le nombre de personnes qui vivent dans les ménages non abonnés qui n'ont eu aucun service l'année précédente et t_4 , le nombre de personnes qui vivent dans les ménages abonnés irrégulièrement qui ne reçoivent pas le service au moment de l'enquête. Le modèle de probabilité de réponse utilisé par ces auteurs se fonde sur l'hypothèse que $t_3 = 0$. Dans ces conditions, un ajustement non biaisé des coefficients de pondération est donné par $A = (t_2 + t_4)/t_2 = 1 + (t_4/t_2)$, c'est-à-dire l'inverse de la proportion de ménages abonnés irrégulièrement qui obtiennent le service au moment de l'enquête. Malheureusement, ces proportions sont inconnues et doivent être estimées. À l'instar de Brick et coll., nous utilisons les données de la CPS pour estimer $t_1 + t_2$, c'est-à-dire le nombre de personnes qui reçoivent le service au moment de l'enquête, ainsi que t_4 , représentons ces estimations par $t_1' + t_2'$ et t_4' , respectivement. D'après l'Appalachian Poll, nous pouvons obtenir des estimations distinctes de t_1 et t_2 que nous représentons par t_1' et t_2' , respectivement. Puisque les estimations proviennent d'enquêtes différentes, nous utilisons des quotients pour l'ajustement des coefficients de pondération et estimons A comme suit

$$A' = 1 + \frac{\frac{\hat{t}_2}{t_2' + t_2'}}{\frac{\hat{t}_4}{t_4' + t_4'}}.$$

(1)

Comme certaines personnes sont plus susceptibles que d'autres de vivre dans un ménage non abonné au téléphone, Brick et coll. répartissent les abonnés irréguliers entre des cellules définies d'après les caractéristiques liées au fait de ne pas avoir le téléphone et calculent la correction de la pondération pour chaque cellule. Ils ont considéré ainsi quatre scénarios de classification, où les répondants sont catégorisés selon le niveau de scolarité ou le mode d'occupation du logement, la durée de l'interruption du service et la race/appartenance ethnique.

Brick et coll. ont conclu que les scénarios consistant à classer les répondants comme des abonnés irréguliers s'ils avaient connu une interruption de service d'au moins une semaine donnent de meilleurs résultats que les scénarios pour lesquels le seuil d'exclusion est une interruption d'une durée d'un mois, si bien que pour les données de l'Appalachian Poll, nous utilisons le seuil d'une semaine. Étant donné le petit nombre d'Hispaniques dans l'échantillon de l'Appalachian-Poll, nous ne procédons pas à la classification selon le groupe ethnique. Par conséquent, pour notre analyse, les classifications en cellules pour les deux scénarios fondés sur la méthode décrite par Brick et coll. (1996) sont les suivantes :

BWKE – ménages qui ont connu une interruption de service d'au moins une semaine dans les catégories définies selon le niveau de scolarité (pas de diplôme d'études secondaires, diplôme d'études secondaires, et selon la race (noire, non noire);

BWKT – ménages qui ont connu une interruption de service d'au moins une semaine dans les catégories définies selon le mode d'occupation du logement (propriétaire/autre, locataire) et la race.

L'inconvénient de l'utilisation de ces scénarios dans notre étude tient au fait que les estimations nécessaires fondées sur la CPS sont disponibles selon l'état, mais non selon la région, car les comtés ne sont pas tous échantillonnés pour réaliser cette enquête. Les habitants des régions appalachiennes sont moins susceptibles d'avoir le téléphone, mais nous ne pouvons rendre compte de ce fait au moyen des données existantes de la CPS. Même si nous considérons des données à l'échelle de l'état, la taille de l'échantillon de la CPS n'est pas suffisante pour produire des valeurs fiables de t_4 pour toutes les cellules. Par exemple, en 1999, l'échantillon de la CPS ne contenait aucun Noir titulaire d'un diplôme collégial ou de niveau supérieur vivant au Kentucky et n'étant pas abonné au

Comme les abonnés irréguliers représentaient une part non négligeable de la population de non-abonnés et que leurs caractéristiques sont plus comparables à celles de cette population qu'à celles des ménages abonnés continuellement, il est raisonnable d'utiliser dans l'échantillon les données recueillies sur les abonnés irréguliers pour essayer de réduire le biais de couverture.

Dans l'Appalachian Poll, 140 des 2463 participants, soit 5,7 %, ont répondu affirmativement à la question « Au cours des 12 derniers mois, votre ménage s'est-il jamais trouvé sans service téléphonique pendant une semaine ou plus? ». Ces répondants sont considérés comme étant des abonnés irréguliers. Le taux d'abonnements irréguliers se chiffre à 7,4 % dans les régions appalachiennes, mais à seulement 3,9 % dans les régions non appalachiennes.

Le tableau 1 donne une comparaison des ménages abonnés irrégulièrement et régulièrement compris dans l'échantillon, pour certaines variables. Les différences prononcées entre les deux populations montrent à quel point il est important de réduire le biais de couverture. Les membres des ménages abonnés irrégulièrement sont nettement plus jeunes, ont un revenu plus faible et sont moins susceptibles d'être occupés à temps plein que ceux des ménages abonnés régulièrement. Ils ont aussi un accès plus limité à l'assurance-maladie et aux ordinateurs.

Tableau 1
Certaines caractéristiques des ménages abonnés régulièrement et irrégulièrement

Caractéristiques		Abonnés réguliers		Abonnés irréguliers	
Âge médian	47,0	37,5	60,0%	27,8%	47,0
Revenu du ménage inférieur à 20 000 \$	55,0%	34,5%	12,7%	30,0%	61,4%
Occupé à temps plein ou retraité	79,4%	47,4%	12,3%	26,4%	42,9%
Propriétaire ou en train d'acheter son logement					
Ordinateur à la maison					
Pas suffisamment d'argent pour se nourrir					

Nota : Étant fondées sur les fréquences non pondérées dans l'échantillon – dans lequel les régions appalachiennes sont surreprésentées – les statistiques ne sont pas représentatives des chiffres de population.

Un modèle d'abonnement irrégulier. Partant de l'échantillon de l'Appalachian Poll, nous développons un modèle de régression logistique pour prédire l'abonnement irrégulier d'après les variables démographiques. Les variables indépendantes utilisées pour prédire l'abonnement irrégulier sont l'âge, la situation d'emploi, la race, le revenu et la région. Le modèle est décrit à l'annexe. Le niveau de escolarité et le mode d'occupation du logement sont aussi de bons prédicteurs de l'abonnement irrégulier, mais, comme

ils sont fortement corrélés à d'autres variables du modèle, nous avons choisi de ne pas les y inclure. Pour une comparaison des modèles qui prédisent la couverture du service téléphonique, consulter Smith (1990). Nous utiliserons notre modèle pour procéder à l'ajustement de la pondération d'après les scores de propension décrit à la section suivante.

3. AJUSTEMENTS DES COEFFICIENTS DE PONDERATION

Nous considérons plusieurs scénarios de pondération qui visent à tenir compte du biais de couverture inhérent aux enquêtes téléphoniques et nous comparons chacun d'eux à la méthode de pondération effectivement utilisée pour l'Appalachian Poll. Dans le cas de la méthode type, nous calculons un coefficient de pondération de base pour chaque répondant. Cet ajustement correspond à (nombre d'adultes dans le ménage)/(nombre de lignes téléphoniques vocales), c'est-à-dire l'inverse de la probabilité que le répondant fasse partie de l'échantillon. Puis, dans chacune des six strates, nous traitons les coefficients de pondération par la méthode d'ajustement proportionnel itératif (*raking ratio*) de sorte que soit respectées les proportions du Recensement de 1990 pour le groupe d'âge, le niveau de escolarité et le sexe. Enfin, nous rééchelonnons les coefficients de pondération en fonction de la taille de l'échantillon dans chacune des six strates.

3.1 Durée de l'interruption du service

Aux participants à l'Appalachian Poll qui ont répondu « oui » à la question concernant l'interruption du service téléphonique pendant au moins une semaine, on a demandé ensuite pendant combien de jours ils n'ont pas été abonnés l'année précédente. Un moyen simple de résoudre le problème du biais de couverture consiste à appliquer aux abonnés irréguliers un ajustement du coefficient de pondération inversement proportionnel à la fraction de l'année durant laquelle leur téléphone n'a pas été branché. Par exemple, une personne qui n'a reçu le service que 6 des 12 derniers mois aura un coefficient de pondération de deux, donc sera représentative d'elle-même et d'un autre membre de la population dont le téléphone a été débranché pendant six mois et qui ne reçoit pas le service au moment de l'enquête.

Nous incluons cette approche naïve dans l'analyse aux fins de comparaison à d'autres scénarios. Nous lui donnerons le nom de scénario du nombre de jours (DAY). L'ajustement des coefficients de pondération est calculé selon la formule $365/(365 - \text{nombre de jours sans service})$. Cet ajustement du coefficient de pondération est appliqué après le calcul du coefficient de base décrit plus haut, mais avant l'ajustement proportionnel itératif (*raking*). Bien qu'elle soit logique, cette méthode n'est pas pratique du point de vue du contrôle de la variance. Il est

pour prédire le phénomène d'abandonnement irrégulier au service téléphonique. À la section 3, nous décrivons en détail les diverses méthodes de pondération. À la section 4, nous discutons du compromis entre la réduction du biais et l'augmentation de la variance liées à l'ajustement des coefficients de pondération et nous comparons les scénarios de pondération. Enfin, à la dernière section, nous résumons nos observations.

2. POPULATIONS DE NON-ABONNÉS ET D'ABONNÉS IRRÉGULIERS AU TÉLÉPHONE

Selon la situation d'abandonnement au service téléphonique, nous pouvons classer la population cible d'une enquête téléphonique en quatre groupes, à savoir les ménages abonnés régulièrement au service téléphonique, les ménages abonnés irrégulièrement au service téléphonique, les ménages abonnés irrégulièrement au moment de l'enquête, les ménages abonnés irrégulièrement au moment qui ne reçoivent pas le service au moment de l'enquête et les ménages chroniquement non abonnés. Pour tenir compte du biais de couverture qui entachent les données de l'enquête, nous devons estimer la taille de chacun de ces groupes. Les données de la FCC permettent de dégager les tendances à long terme quant à la taille de la population de non-abonnés au téléphone. Par contre, on n'en sait pas autant sur les variations à court terme de la couverture téléphonique.

Keeter (1995) s'est servi d'enquêtes par panel pour étudier la dynamique de la population d'abonnés irréguliers au téléphone. Lors des cycles de mars 1992 et 1993 de la CPS, il a constaté que 94,1 % de ménages faisant partie de l'échantillon lors des deux cycles étaient abonnés les deux fois, 2,6 % ne l'étaient aucune fois et 3,4 % l'étaient lors d'une entrevue, mais pas de l'autre. En tout, 57 % de répondants qui ont déclaré ne pas être abonnés au téléphone lors de l'une ou l'autre entrevue étaient des abonnés irréguliers. Si les données pouvaient être recueillies dans le temps, un nombre encore plus grand de ménages pourrait être considéré comme des abonnés irréguliers. Keeter conclut qu'à tout le moins, une minorité importante de ménages non abonnés au téléphone avaient fait partie récemment de la population d'abonnés ou étaient prêts à s'y joindre bientôt et que ces abonnés irréguliers représentaient une part mesurable de la population de ménages abonnés au téléphone et fournissent donc des données permettant de caractériser la population de non-abonnés (Keeter 1995, page 201). Il affirme dans le même article que les ménages abonnés irrégulièrement non abonnés qu'à ceux abonnés régulièrement (Keeter 1995, page 209). Cette conclusion est fondée sur des tests formels appliqués à des variables démographiques de la CPS. Les données provenant de la National Survey of America's Families présentée par Brick et coll. (1999) confirment les observations de Keeter.

d'abonnés irréguliers dans l'échantillon pour représenter la population de non-abonnés. Ces auteurs se servent de données provenant de la U.S. Current Population Survey (CPS) pour calculer un ajustement non biaisé des coefficients de pondération pour les abonnés irréguliers qui participent à leur enquête. Frankel, Ezzati-Rice, Wright et Srinath (1998) appliquent aussi cette correction des coefficients de pondération et considèrent deux ajustements similaires. Brick, Flores Cervantes, Wang et Hanthais (1999) et Frankel, Srinath, Battaglia, Hoaglin, Wright et Smith (1999) évaluent ces ajustements au moyen des données provenant d'enquêtes contenant des questions sur le service téléphonique, mais qui ne sont pas entachées d'un biais de couverture téléphonique. Ces auteurs constatent que l'application de facteurs d'ajustement de la pondération fondés sur la situation d'abonné irrégulier permet généralement d'améliorer les estimations.

Le présent article porte sur une autre méthode d'ajustement de la pondération applicable aux abonnés irréguliers. Cette méthode consiste à élaborer un modèle permettant de prédire l'abandonnement irrégulier en se servant de variables démographiques. L'ajustement de la pondération se fonde alors sur la propension du participant à l'enquête à être abonné irrégulièrement au service téléphonique. Nous comparons aussi cette méthode basée sur les scores de propension à celle proposée par Brick et coll. (1996), ainsi qu'à une méthode fondée sur la probabilité de réponse, où l'ajustement de la pondération se fonde sur la durée de l'interruption du service téléphonique.

Nous nous servons des données de l'Appalachian Poll, enquête téléphonique à composition aléatoire (CA) réalisée en juin et en juillet 1989 par le Center for Survey Research de l'Ohio State University. L'enquête, qui était parrainée par *The Columbus Dispatch*, visait à comparer les régions défavorisées et non défavorisées du Kentucky, de l'Ohio et de la Virginie occidentale. Elle était conçue en vue de recueillir des renseignements sur la qualité de la vie et sur les perceptions au sujet des régions appalachiennes comportant une série de questions types sur les caractéristiques démographiques. Elle a été réalisée auprès d'un échantillon stratifié et un peu plus de 400 questionnaires ont été remplis pour chacune des six strates (régions appalachienne et non-appalachienne de l'Ohio, du Kentucky et de la Virginie occidentale). Le sondage avait pour cible les résidents anglophones de 18 ans et plus des trois états. La question du biais de couverture est particulièrement importante dans le cas de cette enquête, car les taux de pénétration du téléphone sont plus faibles que la normale dans les régions économiquement défavorisées des Appalaches.

À la section 2, nous passons en revue les données publiées sur les populations d'abonnés réguliers et irréguliers au téléphone. Nous y examinons aussi les différences dégagées de nos données entre ces deux groupes, afin de montrer que le biais de couverture est une question préoccupante. Nous concluons la section 2 par notre modèle

Utilisation de scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques

KRISTIN BLENK DUNCAN et ELIZABETH A. STASNY¹

RÉSUMÉ

Les enquêtes téléphoniques représentent une méthode de collecte de données. Cependant, le fait que les ménages n'ayant pas le téléphone en soient exclus peut biaiser les estimations de population. Selon les données de la U.S. Federal Communications Commission (FCC), à tout moment, de cinq et demi à six pour cent de ménages américains ne sont pas abonnés au service téléphonique. Le biais introduit peut être important, car les ménages non abonnés peuvent différer de ceux qui sont abonnés d'une façon dont il n'est pas possible de tenir compte adéquatement par post-stratification. Durant l'année, nombre de ménages, qualifiés d'« abonnés irréguliers », entrent dans la population abonnée au téléphone ou en sortent, parfois pour des raisons économiques ou à cause d'un déménagement. La population d'abonnés irréguliers pourrait être représentative de la population de non-abonnés en général, puisque ses membres ont fait partie récemment de cette deuxième population.

Le présent article décrit l'élaboration d'un ajustement de la pondération tenant compte des abonnés irréguliers en vue de réduire le biais dû au non-dénombrement tout en contrôlant l'augmentation de la variance due à la pondération. Nous utilisons un modèle de régression logistique pour décrire la propension de chaque ménage à être abonné irrégulièrement au service téléphonique, en nous appuyant sur des données provenant d'enquêtes réalisées dans des régions défavorisées et non défavorisées du Kentucky, de l'Ohio et la Virginie occidentale. Les corrections de la pondération se fondent sur les scores de propension. Nous estimons la réduction du biais et de l'erreur d'estimation obtenue pour plusieurs variables d'enquête par application de l'ajustement de la pondération fondé sur les scores de propension et de plusieurs autres méthodes d'ajustement de la pondération. Pour évaluer l'efficacité de l'ajustement, nous comparons l'erreur qui entache les estimations corrigées à celle qui entache les estimations types.

MOTS CLÉS : Enquête par CA; ajustement de la pondération; erreur non due à l'échantillonnage.

1. INTRODUCTION

Étant devenu un mode de communication standard dans notre société, le téléphone s'avère un outil fort utile pour la réalisation d'enquêtes. La fréquence des enquêtes téléphoniques a augmenté parallèlement à la proportion d'abonnés au service téléphonique. Aujourd'hui, la plupart des personnes qui font partie d'une population au sujet de laquelle une enquête vise à faire des inférences, c'est-à-dire la population cible, peuvent être rejointes par téléphone. Par conséquent, on tire l'échantillon de l'ensemble de membres des ménages pouvant être rejoints par composition d'un numéro de téléphone résidentiel. Cependant, cette base de sondage exclut toutes les personnes non abonnées au service téléphonique qui pourraient représenter une part importante de certaines populations. À l'heure actuelle, on estime qu'aux États-Unis, à tout moment, de 5,5 % à 6,0 % des ménages ne sont pas raccordés au service téléphonique (Belinfante 2000). Or, les personnes non abonnées ont tendance à différer de celles qui le sont, particulièrement en ce qui concerne la situation économique (Smith 1990). Les résultats de l'enquête ne seront donc pas fidèlement représentatifs de l'ensemble de la population si ces différences sont significatives pour des caractéristiques importantes du point de vue de l'enquête. Le biais de couverture est particulièrement gênant dans le cas d'enquête visant des

sous-groupes de la population où le taux de pénétration du téléphone est faible. Ces groupes incluent les membres des ménages à faible revenu et les personnes qui n'ont pas terminé leurs études secondaires.

La post-stratification en fonction de variables démographiques associées à la pénétration du téléphone réduit le biais de couverture, mais ne permet pas de résoudre entièrement le problème (Massey et Botman 1988). Notre moyen de tenir compte de ce biais de couverture consiste à permettre que les personnes participant à l'enquête qui, récemment, n'étaient pas abonnées au téléphone représentent celles qui, au moment de l'enquête, ne sont pas abonnées au service téléphonique. Les personnes dont la situation d'abonnement au téléphone a changé durant l'année qui a précédé l'enquête sont considérées comme des abonnés irréguliers. Ces derniers sont des personnes qui entrent dans la population abonnée au téléphone et en sortent, peut-être pour des raisons économiques, ou à cause de l'interruption du service durant un déménagement. Les abonnés irréguliers qui ont un service téléphonique au moment de l'enquête pourraient représenter une partie de la population de non-abonnés, car ils figurent dans la base de sondage mais ont fait partie récemment de la population de non-abonnés.

Brick, Wakseberg et Keeter (1996) ont proposé un ajustement de la pondération fondé sur l'inclusion

¹ Kristin Blenk Duncan et Elizabeth A. Stasny, Department of Statistics, Ohio State University, Columbus, OH 43210-1247.

Marker étudie des stratégies de conception d'enquêtes en vue d'améliorer la qualité des estimateurs régionaux directs et de réduire ainsi la nécessité des estimateurs indirects fondés sur des modèles. Parmi les facteurs pris en considération, mentionnons la stratification et le suréchantillonnage, la combinaison de données provenant d'enquêtes répétées, l'harmonisation d'enquêtes différentes, l'utilisation d'échantillons supplémentaires et le recours à des méthodes d'estimation améliorées.

Dans leur article, Saigo, Shao et Sitter se penchent sur l'importante question de l'estimation de la variance en présence d'imputation pour les données manquantes. Ces auteurs proposent une méthode bootstrap qui fonctionne tant pour des statistiques lissées que pour les statistiques non lissées, même lorsque le nombre de grappes échantillonnées est faible. Cette méthode représente une amélioration par rapport à la méthode bootstrap proposée antérieurement, qui pouvait comporter une importante surestimation dans le cas d'un faible nombre de grappes échantillonnées. Outre cette méthode bootstrap, Saigo, Shao et Sitter proposent une méthode BRR qui prend en compte la variance de l'imputation dans le cas d'une imputation aléatoire. L'utilisation de ces méthodes est démontrée au moyen d'une étude par simulation.

Belhousse et Statford examinent la régression polynomiale locale non paramétrique comme outil exploratoire d'analyse de données à utiliser dans le cas de données provenant d'enquêtes complexes. Ces auteurs considèrent une seule variable explicative continue x , à laquelle on attribue un nombre fini de valeurs possibles pouvant correspondre à l'exactitude de la mesure de x , mais pouvant également être choisies autrement. Par cette méthode, on détermine des estimations ponctuelles de la fonction de régression locale et des estimations de variance correspondantes. La méthode proposée est démontree au moyen d'une analyse d'indices de masse corporelle tirée de l'Enquête sur la santé en Ontario, et les estimations non paramétriques sont comparées aux estimations obtenues à l'aide d'un modèle paramétrique.

Dans le dernier article de ce numéro, Silva et Smith utilisent une méthode fondée sur l'espace et obtenues à l'aide d'un modèle paramétrique. La transformation logarithmique additive et modélisent ensuite la série transformée. On élabore des méthodes d'estimation fondées sur le filtre de Kalman qui sont appliquées à des données tirées de l'Enquête sur la population active du Brésil. Le filtre de Kalman permet également d'obtenir des estimations de la variance fondées sur un modèle ainsi que les limites de confiance pour la série transformée. Les estimations de tendances et les effets saisonniers sont comparés aux valeurs correspondantes obtenues à l'aide de la méthode ARMMI X-11, et généralement, les valeurs obtenues à l'aide de la méthode proposée sont plus lisses, étant donné qu'elles rendent compte explicitement des erreurs d'échantillonnage qui entachent les estimations brutes de la série.

M.P. Singh

Dans ce numéro

Ce numéro de techniques d'enquête renferme des articles sur des sujets variés tels que la couverture, la non-réponse, l'imputation, les plans de sondage, la pondération et l'analyse de données d'enquêtes complexes.

Dans le premier article du présent numéro, Blenk et Stasny proposent une correction de pondération afin de réduire le biais de couverture dans les enquêtes téléphoniques, tout en tenant compte de l'augmentation de la variance due à la pondération. La correction de pondération est appliquée aux ménages de passage, c'est-à-dire aux ménages qui quittent la population des enquêtes téléphoniques no qui intègrent celle-ci au cours de l'année. On suppose que la population transitoire des enquêtes téléphoniques est représentative de la population qui ne participe pas aux enquêtes téléphoniques. La correction de pondération qui est proposée est fondée sur les résultats relatifs à la proposition des ménages à être de passage, qui sont obtenus à l'aide d'un modèle de régression logistique. La méthode proposée ainsi que plusieurs autres méthodes sont comparées entre elles au moyen de données recueillies dans le cadre d'une enquête portant sur des régions en difficulté et des régions non en difficulté des États du Kentucky, de l'Ohio et de Virginie occidentale.

Martiano et Kadane utilisent le nombre d'appels effectués au cours d'une enquête téléphonique comme indicateur du degré de difficulté qu'on éprouve à joindre un répondant donné. Cet indicateur permet une répartition probabiliste des non-répondants, dans un modèle de non-réponse, entre ceux qui vont toujours refuser de répondre et ceux qui ne sont pas disponibles. Il permet également de déterminer si une non-réponse peut être ignorée, pour procéder par inférence à partir de la variable dépendante, en intégrant au modèle l'information sur le nombre d'appels. Ces idées sont appliquées à des données tirées d'une enquête menée au sein de la communauté urbaine de Toronto et portant sur les attitudes vis-à-vis du tabagisme en milieu de travail. Les résultats montrent que la non-réponse n'est pas ignorable et que les non-répondants sont deux fois plus susceptibles d'être en faveur du tabagisme sans restrictions sur les lieux de travail que les personnes qui acceptent de répondre.

Dans son article, Hidiroglou unitifie les cas imbriqués et non-imbriqués que l'on retrouve dans la théorie du double échantillonnage. Le cas imbriqué, aussi appelé échantillonnage à deux phases, correspond au cas classique où l'on tire d'abord un échantillon de première phase permettant de recueillir de l'information auxiliaire et ensuite un échantillon de deuxième phase à l'intérieur du premier échantillon contenant les variables d'intérêt. Le cas non-imbriqué correspond au cas où les deux échantillons sont sélectionnés indépendamment de la même base de sondage ou même de bases de sondage différentes. Un estimateur par la différence généralisée qui peut être utilisé dans les deux cas est proposé et on développe l'estimateur optimal qui minimise la variance. On discute également de l'estimation de la variance pour chacun des cas. Plusieurs exemples d'enquêtes à Statistique Canada illustrent l'unification de ces deux cas.

La Vallée et Caron examinent les difficultés que présente la production de données estimatives lorsqu'on utilise des méthodes de couplage pour lier deux populations entre elles. Ces auteurs se penchent notamment sur les difficultés liées à la production de données estimatives relatives à l'une des deux populations à l'aide d'un échantillon de l'autre population, en supposant que les deux populations ont été mises en liaison. La méthode généralisée du partage des poids est adaptée pour tenir compte des poids de couplage de trois façons : 1) tous les liens sont pris en compte lorsque le poids de couplage est non-zéro; 2) tous les liens sont pris en compte lorsque les poids de couplage sont égaux; 3) on choisit les liens au hasard. Ces estimateurs proposés sont comparés à la méthode classique au moyen d'une étude de simulation.

Merkouris analyse le problème que constitue la production d'estimations transversales à partir de données recueillies dans le cadre d'enquêtes par panels multiples. La couverture de la population transversale peut être incomplète en raison du départ ou de l'arrivée de répondants après la sélection d'un panel. En recomposant qu'une enquête par panel répétitif constitue un type particulier d'enquête à bases de sondage multiples, Merkouris est en mesure de proposer des stratégies de pondération qui conviennent à diverses enquêtes par panels multiples. Ces méthodes de pondération peuvent être utilisées pour agréger l'information tirée de plusieurs panels pour produire des estimations transversales qui tiennent compte de la nature dynamique du plan d'enquête à panels multiples.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 27, numéro 2, décembre 2001

TABLE DES MATIÈRES

Dans ce numéro	129
K. BLENK DUNCAN et E.A. STASNY	
Utilisation de scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques	131
L.T. MARIANO et J.B. KADANE	
Effet de l'intensité des efforts en vue de joindre les répondants : Enquête torontoise sur le tabagisme	143
M.A. HIDIROGLOU	
L'échantillonnage double	157
P. LAVALLÉE et P. CARON	
Estimation par la méthode généralisée du partage des poids : Le cas du couplage d'enregistrements	171
T. MERKOURIS	
Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples	189
D.A. MARKER	
Production d'estimations régionales d'après les données d'enquêtes nationales : Méthodes visant à réduire au minimum l'emploi d'estimateurs indirects	201
H. SAIGO, J. SHAO et R.R. SITTER	
Boostrap à demi-échantillon répété et répliques équilibrées répétées en cas d'imputation aléatoire de données	209
D.R. BELLHOUSE et J.E. STAFFORD	
Régression polynomiale locale dans le cas des enquêtes complexes	219
D.B.N. SILVA et T.M.F. SMITH	
Modélisation de séries chronologiques compositionnelles d'après des données d'enquêtes répétées	227
Remerciements	239

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président	G.J. Brackstone
Membres	D.A. Binder G.J.C. Hole E. Rancourt (Gestionnaire de la production) C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, Statistique Canada

Rédacteurs associés

D.R. Bellhouse, University of Western Ontario
D.A. Binder, Statistique Canada
J.M. Brick, Westat, Inc.
C. Clark, U.S. Bureau of the Census
J.-C. Deville, INSEE
J. Ellings, U.S. Bureau of Labor Statistics
W.A. Fuller, Iowa State University
J. Gambino, Statistique Canada
M.A. Hidiroglou, Statistique Canada
D. Holt, University of Southampton, U.K.
G. Kalton, Westat, Inc.
P. Kott, National Agricultural Statistics Service
P. Lahiri, Joint Program in Survey Methodology
S. Linacre, Official National Statistics

Rédacteurs adjoints

G. Nafhan, Hebrew University, Israel
D. Norris, Statistique Canada
D. Pfeffermann, Hebrew University
J.N.K. Rao, Carleton University
T.J. Rao, Indian Statistical Institute
L.-P. Rivest, Université Laval
F.J. Schuren, National Opinion Research Center
R. Sitter, Simon Fraser University
C.J. Skinner, University of Southampton
E. Stasny, Ohio State University
R. Valliant, Westat, Inc.
J. Waksberg, Westat, Inc.
K.M. Wolter, National Opinion Research Center
A. Zaslavsky, Harvard University

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ CA par année. Le prix n'inclut pas les taxes de ventes canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires, autres pays, 20 \$ CA (10 \$ x 2 exemplaires). Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division de la diffusion, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au 1 800 700-1033, par télécopieur au 1 800 889-9734 ou par Courriel: order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale des Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiennes et statisticiens du Québec.

Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Février 2002

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

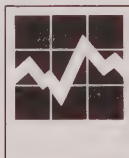
© Ministre de l'Industrie, 2002

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 2001 • VOLUME 27 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



164



NUMÉRO 2

•

VOLUME 27

•

DÉCEMBRE 2001

PAR STATISTIQUE CANADA

ÉDITÉE

UNE REVUE

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE

